Molecular Plant Pathology

MOLECULAR PLANT PATHOLOGY (2018) **19**(9), 2094–2110 DOI: 10.1111/mpp.12682

Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0

JANA SPERSCHNEIDER[1,*], PETER N. DODDS[2], DONALD M. GARDINER[3], KARAM B. SINGH[1,4] AND JENNIFER M. TAYLOR[2]

[1] *Centre for Environment and Life Sciences, CSIRO Agriculture and Food, Perth, WA 6014, Australia*
[2] *Black Mountain Laboratories, CSIRO Agriculture and Food, Canberra, ACT 2601, Australia*
[3] *CSIRO Agriculture and Food, Queensland Bioscience Precinct, Brisbane, Qld 4067, Australia*
[4] *Department of Environment and Agriculture, Centre for Crop and Disease Management, Curtin University, Bentley, WA 6102, Australia*

SUMMARY

Plant-pathogenic fungi secrete effector proteins to facilitate infection. We describe extensive improvements to EffectorP, the first machine learning classifier for fungal effector prediction. EffectorP 2.0 is now trained on a larger set of effectors and utilizes a different approach based on an ensemble of classifiers trained on different subsets of negative data, offering different views on classification. EffectorP 2.0 achieves an accuracy of 89%, compared with 82% for EffectorP 1.0 and 59.8% for a small size classifier. Important features for effector prediction appear to be protein size, protein net charge as well as the amino acids serine and cysteine. EffectorP 2.0 decreases the number of predicted effectors in secretomes of fungal plant symbionts and saprophytes by 40% when compared with EffectorP 1.0. However, EffectorP 1.0 retains value, and combining EffectorP 1.0 and 2.0 results in a stringent classifier with a low false positive rate of 9%. EffectorP 2.0 predicts significant enrichments of effectors in 12 of 13 sets of infection-induced proteins from diverse fungal pathogens, whereas a small cysteine-rich classifier detects enrichment in only seven of 13. EffectorP 2.0 will fast track the prioritization of high-confidence effector candidates for functional validation and aid in improving our understanding of effector biology. EffectorP 2.0 is available at http://effectorp.csiro.au.

Keywords: effector, EffectorP, effector prediction, fungal pathogens, machine learning, secretomes.

INTRODUCTION

Fungal pathogens have been estimated to cause annual crop yield losses of 15%–20% and are a major threat to food security (Figueroa *et al*., 2007; Fischer *et al*., 2012). Fungi colonize plants through diverse infection structures and the use of toxic fungal secondary metabolites and secreted effector proteins that alter host cell structure and function, suppress plant defence responses or modulate plant cell physiology (Kamoun, 1983; Lo Presti *et al*., 2014). Effectors are used by plant-pathogenic fungi and symbiotic fungi to allow them to colonize their hosts. Fungal effectors can be attached to the fungal cell wall, can function in the plant apoplast or can translocate into plant cells where they may target specific host proteins or enter subcellular compartments (Lo Presti *et al*., 2014). Accurate effector mining from genomic sequences is crucial to subsequent experimental validation and effector identification can enable disease control strategies. For example, effectors can be used directly in effector-assisted breeding to select plant lines with distinct recognition traits (Vleeshouwers and Oliver, 2014), and the identification of both effectors and their targets could allow 'decoy engineering', where effector targets are fused as baits to a plant immune receptor to make an integrated 'effector trap' (Ellis, 2011).

Recent progress in big data genomics has resulted in many high-quality fungal pathogen genomes and gene expression profiles during plant infection, but accurate effector prediction methods are needed to harness the potential of these resources. The set of secreted proteins expressed during infection is typically too large for experimental investigation and contains many secreted non-effectors that play roles in niche colonization and protection from competing microbes, differentiation of fungal structures and cell-to-cell communication (Rovenich *et al*., 2014). Secreted plant cell wall-degrading enzymes (PCWDEs) are used by saprophytic fungi to acquire sugars for their nutrition and survival (Kubicek *et al*., 2011). Necrotrophic plant-pathogenic fungi use PCWDEs to overcome the barrier of the cell wall, as well as for nutrient acquisition, whereas biotrophic plant-pathogenic fungi utilize PCWDEs to facilitate stealth invasion of living plant cells (Gibson *et al*., 2001).Some PCWDEs in plant-pathogenic fungi may include effectors specifically required for penetration (Lo Presti *et al*., 2014); however, these can be predicted based on the presence of conserved enzymatic structures or sequence domains. In contrast, the vast majority of fungal effectors are diverse in sequence and share no conserved sequence motifs or obvious commonalities, apart from their secretion from pathogen to the host. This lack of apparent unifying sequence-based features has led to *ad-hoc* fungal

[*] *Correspondence*: Email: jana.sperschneider@csiro.au

effector prediction approaches that are based on various combinations of characteristics observed in known effectors, such as a small protein size, a high cysteine content, evidence of diversifying selection, the genomic location of the gene in fast-evolving regions or gene expression *in planta* (Sperschneider *et al*., 2017). The inclusion of only a few features in effector prediction, such as the requirement of a small protein size, typically results in many false positive predictions and often overwhelmingly large effector candidate sets, such as 1088–2092 effector candidates predicted in stripe rust (Petre *et al*., 2011). However, the inclusion of additional features associated with effectors will capture only a small subset as none of these signals are common to all effectors. For example, some fungal effectors are highly enriched in cysteines, whereas others do not feature any cysteines in their sequence, and fungal effectors also vary in size. For example, the *Pyrenophora tritici-repentis* ToxB effector has 87 amino acids with four cysteines, and is thought to function in the plant apoplast (Figueroa *et al*., 2017), whereas the *Melampsora lini* AvrM effector has a sequence length of 314 amino acids and only one cysteine, and acts intracellularly (Catanzariti *et al*., 2004). However, a high cysteine content or small protein size alone does not allow for the accurate discrimination of apoplastic effectors from cytoplasmic effectors in fungi (Sperschneider *et al*., 2008). Taken together, the use of predefined criteria for effector prediction inherits the individual researcher's potentially biased view of effector characteristics and is unable to uncover novel effectors with diverse characteristics.

An alternative approach is to use data to learn which features are important for effector prediction, rather than setting predefined criteria. This is achieved with machine learning, a family of statistical learning methods with the ability to identify patterns in data and recognize a particular class based on its features in observed data. Models trained on datasets of positive and negative classes are then applied to identify new instances of the class in unseen data. This data-driven approach has the capacity to identify new features not apparent to manual inspection and to provide probabilistic predictions based on combinations of features, which represent advantages over the use of predefined criteria with binary cut-offs. We have recently pioneered such a machine learning approach for fungal effector prediction, called EffectorP (Sperschneider *et al*., 1996), and have demonstrated that machine learning can accurately predict novel effectors with diverse characteristics from secretomes, as well as their localization in the plant cell (Sperschneider *et al*., 2011, 2011). We have shown that EffectorP 1.0 is able to learn 'effector rules' from positive and negative training examples without having to apply user-chosen thresholds (Sperschneider *et al*., 1996). EffectorP relies on fungal effectors as the positive training set and secreted non-effectors as the negative set. One limiting factor is that the negative training set consists of both undiscovered effectors and secreted non-effectors, and therefore poses an unlabelled data classification problem. Furthermore, the positive training set used in EffectorP 1.0 is small and additional effectors are now available for inclusion in training. This has the potential to improve accuracy and will enable us to re-evaluate the ability of machine learning to accurately predict fungal effectors.

RESULTS

Training of the ensemble classifier EffectorP 2.0

EffectorP 1.0 is a Naïve Bayes classifier that was trained on a positive training set of 58 experimentally supported fungal effectors from 16 fungal species. Since its development, additional fungal effectors have been described and, for EffectorP 2.0, we used an expanded training set of 94 secreted fungal effectors from 23 species (Table 1). EffectorP 1.0 predicts 73% of the unseen effectors correctly, which demonstrates its ability to identify novel effectors, but also leaves room for improvement. We set out to investigate whether re-training of EffectorP would improve prediction accuracy.

EffectorP 1.0 was trained on a negative set consisting of predicted secreted proteins from the same pathogen species as the known effectors. Thus, the negative training set includes both undiscovered effectors and non-effectors, and therefore poses an unlabelled data classification problem. Although Naïve Bayes classifiers are fairly robust to unlabelled data classification and can tolerate noisy data (Bing *et al*., 2007), other machine learning classifiers might not be able to learn effectively from such sets. To improve predictions, we collected three different subsets of negative training data that are less likely to contain positive instances, i.e. fungal effectors. First, secretomes were predicted from the same fungal pathogen/symbiont species as used in the positive set if they had a publicly available predicted gene set (Table 1). The combined secretome was homology reduced and this resulted in a filtered predicted pathogen secretome of 11 277 proteins. This set will contain both undiscovered effectors and secreted non-effectors, which poses a challenge for machine learning classifiers that traditionally learn from labelled data. Therefore, we applied EffectorP 1.0 to exclude predicted effectors from the secretomes ($n = 6138$). This procedure removed predominantly small, cysteine-rich proteins from the negative training set (average sequence length, 137 amino acids; average cysteine content, 3.55%). We also collected homology-reduced sets of secreted fungal proteins from fungi not pathogenic on plants, namely from 27 saprophyte secretomes ($n = 12\ 939$) and from 10 animal-pathogenic fungal secretomes ($n = 2763$). These sets are less likely to contain plant-pathogenic effectors and were not filtered for EffectorP 1.0-predicted effectors.

As we have large amounts of negative training data ($n = 21\ 840$), we used an ensemble learning approach of

Table 1 The set of fungal effector proteins used as positive training data.

Species	Effector
Melampsora lini	AvrM, AvrL567-A, AvrP123, AvrP4, **AvrM14**, **AvrL2-A**
Uromyces fabae	RTP1
Puccinia graminis f. sp. *tritici*	PGTAUSPE-10-1, **AvrSr50**
Puccinia striiformis f. sp. *tritici*	**PstSCR1**, **Pec6**
Phakopsora pachyrhizi	**PpEC23**
Blumeria graminis f. sp. *hordei*	**Avrk1**, **Avra1**, **Avra13**
Blumeria graminis f. sp. *tritici*	**AvrPm2**
Cladosporium fulvum	Avr9, Avr4, Avr4E, Avr2, Avr5, Ecp1, Ecp2, Ecp4, Ecp5, Ecp6
Leptosphaeria maculans	AvrLm6, AvrLm4–7, AvrLm1, AvrLm11
Fusarium oxysporum f. sp. *lycopersici*	Six4, Six3, Six1, Six6, Six2, Six5, Six7, Six8
Magnaporthe oryzae	Avr-Pita, Pwl1, Avr-Pia, Bas3, Bas2, Bas4, Bas1, MC69, AvrPiz-t, Avr1-CO39, Avr-Pii, Avr-Pik, Bas107, **AvrPib**, **Iug6**, **Iug9**, **Msp1**, **MoHEG13**, **MoCDIP1**, **MoCDIP2**, **MoCDIP3**, **MoCDIP4**, **MoCDIP5**, **SPD2**, **SPD4**, **SPD7**, **SPD9**, **SPD10**, **Bas162**, **AvrPi9**
Rhynchosporium secalis	NIP1, NIP2, NIP3
Verticillium dahliae	Vdlsc1, Ave1, **VdSCP7**, **PevD1**
Ustilago maydis	Cmu1, Pep1, Pit2, Tin2, **eff1-1**, **See1**
Ustilago hordei	UhAvr1
Stagonospora nodorum	ToxA, Tox1, Tox3
Botrytis cinerea	Nep1
Pyrenophora tritici-repentis	ToxB
Laccaria bicolor	MiSSP7
Zymoseptoria tritici	**AvrStb6**, **Zt6**
Colletotrichum graminicola	**CgEP1**, **Cgfl**
Fusarium graminearum	**FGL1**
Sclerotinia sclerotiorum	**SsSSVP1**

Ninety-four fungal effectors were collected from the literature if they had experimental support and did not share sequence homology. Effectors that were not part of the EffectorP 1.0 training set are marked in bold. All sequences are available at: http://effectorp.csiro.au/data.html.

classifiers that each take a different subset of negative training data and thus provide a different view on classification (Fig. 1). Overall, we chose a total of 50 best-performing models comprising: 10 Naïve Bayes classifiers and 10 C4.5 decision trees that discriminate between fungal effectors and secreted pathogen proteins; 10 Naïve Bayes classifiers and 10 C4.5 decision trees that discriminate between fungal effectors and secreted saprophyte proteins; and five Naïve Bayes classifiers and five C4.5 decision trees that discriminate between fungal effectors and secreted animal pathogen proteins. In 10-fold cross-validation, the Naïve Bayes classifiers achieve, on average, high sensitivity, whereas the C4.5 decision trees show high specificity (Table S2, see Supporting Information). To generate EffectorP 2.0, we combined these 50 models into an ensemble classifier to utilize their distinct prediction strengths (Fig. 1). Each model has seen a different subset of negative training data and, for a given protein sequence input, returns a probability of whether it is an effector or a non-effector. EffectorP 2.0 returns a final prediction using a voting approach on the predicted probabilities of each model. A protein is classified as an effector if the average probability for the class ‘effector’ is higher than the average probability for the class ‘non-effector’. For each protein in the training set, EffectorP 1.0 utilizes a feature vector that is calculated using amino acid frequencies, amino acid class frequencies, molecular weight, sequence length and protein net charge (Sperschneider *et al*., 1996). EffectorP 2.0 uses an updated feature vector that includes amino acid frequencies, amino acid class frequencies, molecular weight, protein net charge, grand average of hydrophobicity, as well as the averages of surface exposure, disorder propensity, hydrophobicity, bulkiness and interface propensity (Table 2).

Influential features for effector prediction include protein size, protein net charge as well as the amino acids serine and cysteine

To detect the most discriminative features in the EffectorP 2.0 classification, we analysed the distribution of features for the proteins employed in the training of all 50 models. Four features were found to be different at a significance threshold of $P < 10^{-5}$ in distribution between the positive sequence set (effectors) and the negative sequence set (proteins labelled as non-effectors) (Fig. 2). Differences in feature distribution for these four features were also reported previously in the EffectorP 1.0 model as particularly striking (Sperschneider *et al*., 1996), confirming their importance in fungal effector classification. As a group, the effectors exhibit lower molecular weight, a higher percentage of cysteines (C) and a lower percentage of serines (S) than the proteins in the negative sequence set. The distribution of protein net charge for effectors occupies a narrow range around neutral to slightly

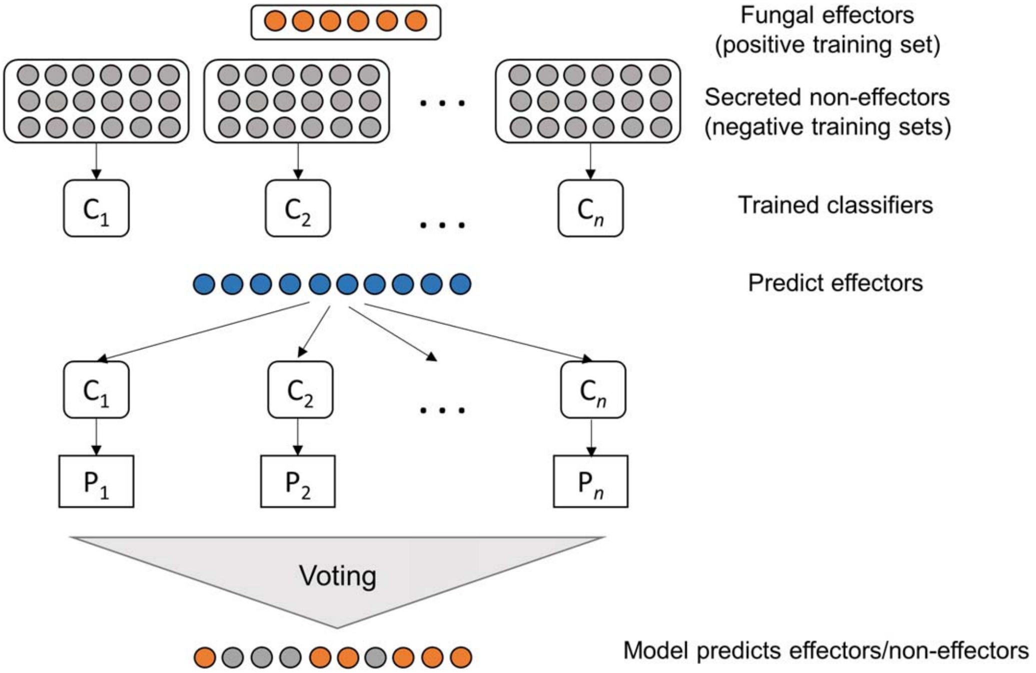

Fig. 1 Workflow for the EffectorP 2.0 classifier that combines an ensemble of machine learning classifiers. Each classifier C_i has seen a different subset of the negative training data and predicts effectors in unseen data with probability P_i. The probabilities are combined into an overall vote on whether an unseen protein is an effector or non-effector.

positive (Fig. 2). We also found significant differences ($P < 0.05$) in distribution between effectors and the negative sequence set for additional features (Fig. 2). These were depletion in aliphatic amino acids, leucine (L), proline (P), threonine (T), tryptophan (W), disorder propensity and bulkiness, as well as enrichment in basic amino acids, interface propensity, glycine (G), lysine (K) and asparagine (N), for effectors. Only enrichment in tryptophan content in effectors was also reported in the EffectorP 1.0 model.

Table 2 Features used for training the machine learning classifiers in the EffectorP 2.0 ensemble learner.

Features used in training and classification	Method
Frequencies of amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) in the sequence	pepstats (Rice *et al.*, 2000)
Frequencies of amino acid classes in the sequence:	
Tiny (A+C+G+S+T)	
Small (A+B+C+D+G+N+P+S+T+V)	
Aliphatic (I+L+V)	
Aromatic (F+H+W+Y)	
Non-polar (A+C+F+G+I+L+M+P+V+W+Y)	
Polar (D+E+H+K+N+Q+R+S+T+Z)	
Charged (B+D+E+H+K+R+Z)	
Basic (H+K+R)	
Acidic (B+D+E+Z)	
Molecular weight	
Protein net charge	
Grand average of hydropathicity (GRAVY, Kyle and Doolittle, 1982)	ProtParam (Gasteiger *et al.*, 2005)
Average of surface exposure (Janin, 1979)	Amino acid groupings and scales taken from Composition Profiler (Vacic *et al.*, 2007)
Average of disorder propensity (Dunker *et al.*, 2001)	
Average of hydrophobicity (Fauchere and Pliska, 1983)	
Average of bulkiness (Zimmerman *et al.*, 1968)	
Average of interface propensity (Jones and Thornton, 1997)	

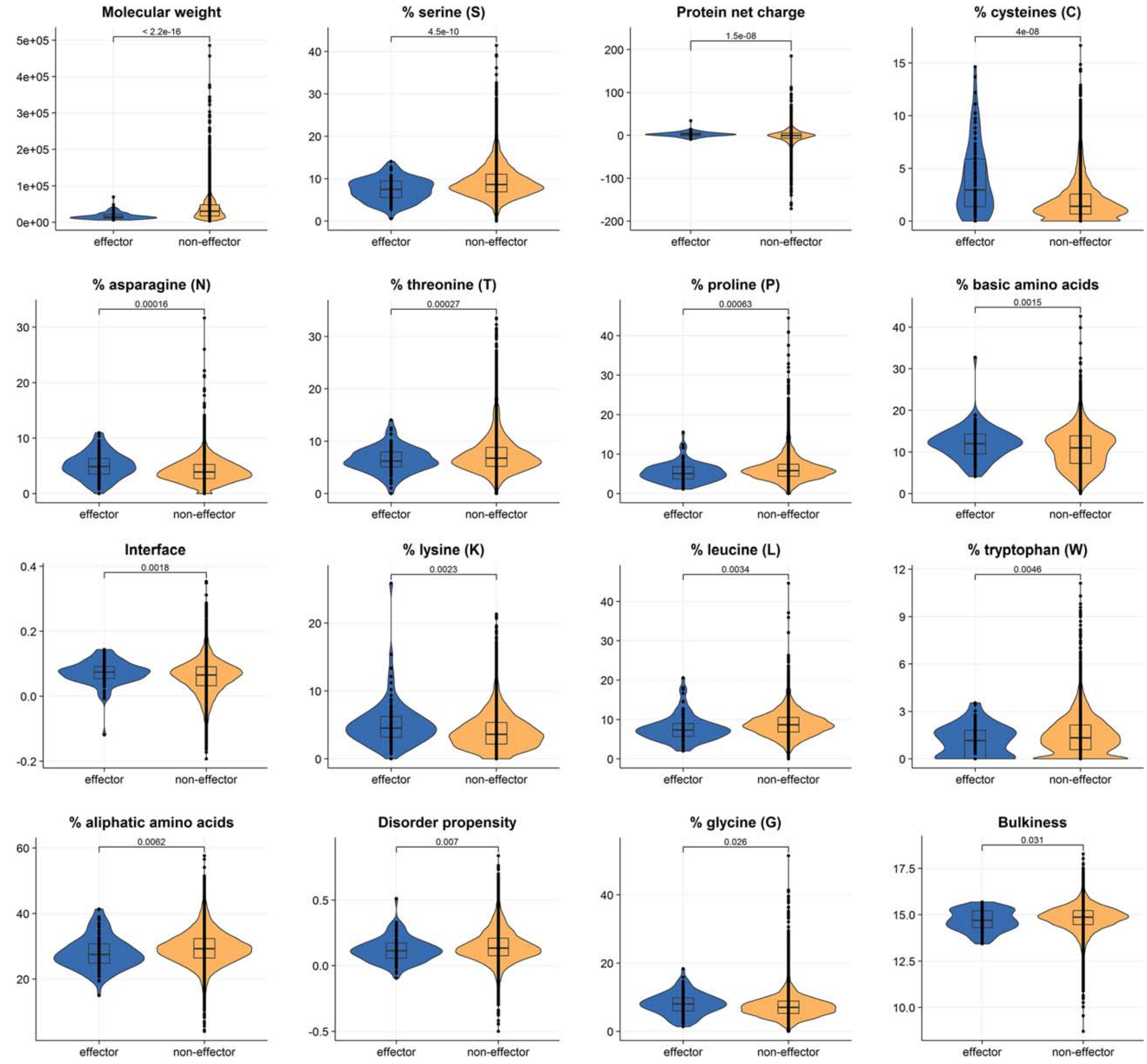

Fig. 2 The most influential features in effector prediction appear to be a small protein size, low serine content, a protein net charge around the neutral range and a high cysteine content. Significant differences ($P < 0.05$) in distribution between effectors and the negative sequence set for additional features were also observed. These were depletion in aliphatic amino acids, leucine (L), proline (P), threonine (T), tryptophan (W), disorder propensity and bulkiness, as well as enrichment in basic amino acids, interface propensity, glycine (G), lysine (K) and asparagine (N), for effectors. Extreme outliers in the protein net charge plot were removed for clarity (full figure given in Fig. S3, see Supporting Information). All data points are drawn on top of the box plots as black dots. Significance between groups is shown as horizontal brackets and was assessed using *t*-tests. The lower and upper hinges correspond to the first and third quartiles and the upper (lower) whiskers extend from the hinge to the largest (smallest) value that is within 1.5 times the interquartile range of the hinge. Data beyond the end of the whiskers are outliers.

Machine learning can be a black box learning process where the reasons for an individual prediction are hidden. However, C4.5 decision trees are white box models and their decision-making process is transparent through navigation along tree branches. As examples, we plotted two of the 10 C4.5 decision trees that discriminate between fungal effectors and secreted pathogen proteins (Figs S1 and S2, see Supporting Information). This demonstrates that the decision tree classifiers use a complex set of features and not only the most discriminative features (protein size, protein net charge as well as the amino acids serine and cysteine) for effector classification. In particular, the decision tree in Fig. S2 does not utilize serine content as a feature in classification and still achieves high classification accuracy. Taken together, this analysis confirms the importance of specific combinations of

Table 3 Independent validation of EffectorP's prediction accuracy.

		Predicted effectors				
Dataset	# of proteins	EffectorP 2.0	EffectorP 1.0	EffectorP 1.0 and 2.0	Small size classifier	Small, cysteine-rich classifier
Fungal saprophyte secreted proteins	24 432	2865 (11.7%)	4774 (19.5%)	2444 (10%)	10 529 (43.1%)	4961 (20.3%)
Fungal, plant and mammalian proteins with signal peptide and localization to endoplasmic reticulum, Golgi, membranes or with glycosylphosphatidylinositol (GPI) anchors	2631	220 (8.4%)	294 (11.2%)	164 (6.2%)	654 (24.9%)	307 (11.7%)
Fungal proteins with unaffected pathogenicity phenotype	938	45 (4.8%)	59 (6.3%)	36 (3.8%)	128 (13.6%)	60 (6.4%)
	28 001	**3130 (11.2%)**	**5127 (18.3%)**	**2644 (9.4%)**	**11 311 (40.4%)**	**5328 (19%)**
Fungal effector positive training set	94	89 (94.7%)	80 (85.1%)	79 (84%)	88 (93.6%)	53 (56.4%)
Fungal effector independent test set	21	16 (76.2%)	16 (76.2%)	16 (76.2%)	19 (90.5%)	10 (47.6%)
Accuracy		88.8%	81.7%	**90.5%**	59.8%	80.9%

features, as found previously in the EffectorP 1.0 model, but also illustrates that accurate fungal effector prediction machine learning classifiers rely on a diverse set of features.

EffectorP 2.0 improves fungal effector prediction accuracy from secretomes

Machine learning classifiers can overfit/overtrain to memorize the training data, which leads to low accuracy on unseen data. Therefore, independent test sets are important to estimate prediction ability. We collected independent positive and negative test sets to assess the performance of EffectorP 2.0. To estimate the false positive rate, we first used fungal, plant and mammalian proteins with predicted signal peptides that were not extracellular [localization to endoplasmic reticulum, Golgi or membranes or with glycosylphosphatidylinositol (GPI) anchors]. A low false positive rate on these proteins ensures that EffectorP is not merely predicting the presence of a signal peptide. We also used secreted saprophyte proteins as well as fungal proteins from PHI-base (Urban *et al*., 2007) that were annotated as having an unaffected pathogenicity phenotype. Although proteins with an unaffected pathogenicity phenotype are not necessarily non-effectors, we expect to see a low percentage of predicted effectors. A simple classifier based on a small protein size ($\leq$300 amino acids) has a false positive rate of 40.4% on these three sets. A small, cysteine-rich classifier ($\leq$300 amino acids; $\geq$4 cysteines) has a false positive rate of 19%, and EffectorP 1.0 has a false positive rate of 18.3%. EffectorP 2.0 has the lowest false positive rate of 11.2% (Table 3). A combination of EffectorP 1.0 and 2.0, where a protein is a predicted effector only if both classifiers label it as an effector, achieves the lowest false positive rate of 9.4%.

To assess false negative predictions, we also applied these predictors to the training data of 94 fungal effectors (Table 3). EffectorP 2.0 only predicts five of these proteins as non-effectors: the *Phakopsora pachyrhizi* effector PpEC23, the *Blumeria graminis* f. sp. *hordei* effector Avrk1, the *Magnaporthe oryzae* effector MoCDIP2, the *Ustilago maydis* effector eff1-1 and the *Colletotrichum graminicola* metalloproteinase effector Cgfl. This is an improvement on EffectorP 1.0, which correctly predicted only 80 of the 94 positive examples. However, it is also important to assess overfitting on training data and to use unseen fungal effectors independent from the training set for the validation of the estimated true positive rate. Therefore, we collected 21 effectors (Table 4) that either shared sequence similarity with an effector in the training set and were therefore eliminated in the homology reduction step (Mg3LysM, BEC1054, BEC1011, AvrLm2) or were overlooked during initial literature searches for training the EffectorP 2.0 model (SAD1, CSEP-07, CSEP-09, SIS1, CSEP0055, BEC1019, Bcg1, CSEP0105, CSEP0162, AvrLmJ1, AvrLm3, XylA, Ecp7, PIIN_08944, FGB1, AvrPm3, AvrSr35). On this independent test set, both EffectorP 1.0 and 2.0 show equal performance and correctly predict 76.2% of effectors (Tables 3 and 4). On the total set of 115 effectors, the small size classifier correctly predicts 93% of effectors, but the small, cysteine-rich classifier only correctly predicts 54.8% of effectors. On the combined positive and negative sets, EffectorP 2.0 has the highest accuracy of 88.8% of the four single classifiers. The simple classifier based on a small size has the lowest accuracy of 59.8%, largely because of its high false positive rate (Table 3). The combined EffectorP 1.0/2.0 classifier achieves the highest accuracy of 90.5% because of its low false positive rate. Although the combined EffectorP 1.0/2.0 classifier misses more effectors than EffectorP 2.0 or 1.0, it is a highly stringent method for the prediction of effectors in secretomes. In the following, we assess the prediction abilities of EffectorP 1.0 compared with EffectorP 2.0 in more detail.

Sets of infection-induced proteins are enriched for effectors predicted by EffectorP 2.0

Effectors are often induced during infection, and thus the set of genes differentially expressed during infection should be enriched

Table 4 Independent test set of fungal effectors that were not used in training of EffectorP 2.0

Species	Effector	EffectorP 1.0 (probability)	EffectorP 2.0 (probability)	Small size classifier	Small, cysteine-rich classifier
Sporisorium reilianum	SAD1	Effector (0.97)	Effector (0.621)	Effector	Non-effector
Phakopsora pachyrhizi	CSEP-07	Effector (0.608)	Effector (0.688)	Effector	Effector
	CSEP-09	Effector (0.999)	Effector (0.842)	Effector	Effector
Zymoseptoria tritici	Mg3LysM	Non-effector (0.556)	Non-effector (0.561)	Effector	Effector
Blumeria graminis f. sp. *hordei*	BEC1054	Effector (0.935)	Effector (0.869)	Effector	Non-effector
	BEC1011	Effector (0.974)	Effector (0.947)	Effector	Non-effector
	BEC1019	Non-effector (0.986)	Non-effector (0.551)	Non-effector	Non-effector
	CSEP0055	Effector (0.649)	Effector (0.732)	Effector	Non-effector
	Bcg1	Effector (0.971)	Effector (0.896)	Effector	Non-effector
	CSEP0105	Non-effector (0.511)	Non-effector (0.595)	Effector	Effector
	CSEP0162	Effector (0.854)	Effector (0.693)	Effector	Effector
Rhizophagus irregularis	SIS1	Effector (0.973)	Effector (0.611)	Effector	Non-effector
Leptosphaeria maculans	AvrLmJ1	Effector (0.999)	Effector (0.727)	Effector	Effector
	AvrLm2	Effector (0.764)	Effector (0.578)	Effector	Effector
	AvrLm3	Effector (1.0)	Effector (0.91)	Effector	Effector
Fusarium graminearum	XylA	Effector (0.882)	Effector (0.865)	Effector	Non-effector
Cladosporium fulvum	Ecp7	Effector (0.997)	Effector (0.96)	Effector	Effector
Piriformospora indica	PIIN_08944	Non-effector (0.886)	Non-effector (0.539)	Effector	Non-effector
	FGB1	Effector (1.0)	Effector (0.929)	Effector	Effector
Blumeria graminis f. sp. *tritici*	AvrPm3	Effector (0.979)	Effector (0.913)	Effector	Non-effector
Puccinia graminis f. sp. *tritici*	AvrSr35	Non-effector (1.0)	Non-effector (0.918)	Non-effector	Non-effector

for effectors. However, not all genes that are differentially expressed during infection encode effector proteins, and therefore sets of differentially expressed genes need to be filtered further to detect effectors. We collected 13 gene sets from the literature that were labelled as containing effector candidates based on their induction during infection as well as other criteria (Table 5). For example, a study by Germain *et al*. (2011) identified 16 candidate effectors from 1184 small, secreted *Melampsora larici-populina* proteins. These 16 candidates were selected based on their expression in a haustoria-specific cDNA library and the transcriptome of laser microdissected, rust-infected poplar leaves, as well as their small size of less than 300 amino acids. As another example, Kettles *et al*. (2017) selected 63 *Zymoseptoria tritici* candidate effectors on the basis of being induced during early wheat leaf infection leading up to the transition to the necrotrophic growth phase. In total, four of the 13 sets contained infection-induced effector candidates that were pre-selected based on a small size (≤300 amino acids).

We assessed whether the 13 sets containing infection-induced effector candidates are also enriched for effector candidates predicted by EffectorP 1.0 or 2.0, by a small size classifier or by a small, cysteine-rich classifier when compared with the whole secretome of each species. We did not test the small size classifier on sets containing effector candidates that were pre-selected based on a small size (≤300 amino acids). We found significant enrichments for predicted effector candidates in 12 of 13 sets (92.3%) using EffectorP 2.0 (Table 5). A small, cysteine-rich classifier only returns significant enrichments for predicted effectors in seven of 13 sets (53.9%) and EffectorP 1.0 in 10 of 13 sets (76.9%). A small size classifier returns significant enrichments for predicted effectors in eight of nine sets (88.9%). Surprisingly, we did not observe enrichment for predicted effectors with any of the four classifiers in secreted proteins of *P. graminis* f. sp. *tritici* highly up-regulated in haustoria compared with germinated spores (Table 5). This could indicate that rusts might utilize undiscovered effector proteins with different properties to the training set, such as effectors of larger size. This is supported by the recent discovery of AvrSr35, a 578-amino-acid *P. graminis* f. sp. *tritici* effector protein (Salcedo *et al*., 2005). Alternatively, haustorial secretomes might contain many non-effectors, such as proteins involved in signalling or in the incorporation of nutrients from the host (Garnica *et al*., 2005). Taken together, although effector function has not been established for all genes in these candidate sets, the enrichment for predicted effectors in infection-induced sets underlines the ability of EffectorP 2.0 to accurately predict unseen effectors.

EffectorP 2.0 reduces the average number of effectors predicted for fungal plant symbionts and saprophytes by 40%

We tested EffectorP 2.0 on predicted secretomes from 93 fungal species, including pathogens and non-pathogens (Table S3, see Supporting Information), and recorded the percentages of secreted proteins that are predicted effectors (Table S4, see Supporting Information). The highest proportions of predicted effectors were found in the obligate biotrophs *Melampsora laricis-populina* (41.3%), *Puccinia graminis* f. sp. *tritici* (40.3%), *Blumeria graminis* f. sp. *hordei* (38.1%) and *Puccinia striiformis* f. sp. *tritici*

Table 5 Enrichment of predicted effector candidates in expression datasets of early infection stages.

Expression dataset	No. of proteins	Method	Predicted effectors	Predicted effectors in secretome	Enrichment (Fisher's exact test)
Colletotrichum higginsianum: biotrophy-associated effector candidates (Kleemann *et al*., 2012)	102	Small size	100 (98%)	845 (56.6%)	<**0.0001**
		Small, cysteine-rich	46 (45.1%)	412 (27.6%)	**0.0003**
		EffectorP 1.0	73 (71.6%)	490 (32.8%)	<**0.0001**
		EffectorP 2.0	49 (48%)	378 (25.3%)	<**0.0001**
Cladosporium fulvum: *in planta* induced small secreted apoplastic effector candidates (Mesarich *et al*., 2017)	75	Small size	–	–	**–**
		Small, cysteine-rich	70 (93.3%)	272 (25%)	<**0.0001**
		EffectorP 1.0	68 (90.7%)	237 (21.8%)	<**0.0001**
		EffectorP 2.0	64 (85.3%)	190 (17.5%)	<**0.0001**
Magnaporthe oryzae: genes with $\geq$50-fold differential expression in biotrophic invasive hyphae (Mosquera *et al*., 2009)	15	Small size	15 (100%)	907 (55.6%)	**0.0002**
		Small, cysteine-rich	9 (60%)	500 (30.7%)	**0.0221**
		EffectorP 1.0	14 (93.3%)	614 (37.7%)	<**0.0001**
		EffectorP 2.0	13 (86.7%)	489 (30%)	<**0.0001**
Blumeria graminis f. sp. *hordei*: Candidates for Secreted Effector Proteins (CSEPs) (Pedersen *et al*., 2012)	491	Small size	347 (70.7%)	426 (58.8%)	<**0.0001**
		Small, cysteine-rich	133 (27.1%)	169 (23.3%)	NS
		EffectorP 1.0	274 (55.8%)	302 (41.7%)	<**0.0001**
		EffectorP 2.0	256 (52.1%)	276 (38.1%)	<**0.0001**
Melampsora larici-populina: specific small secreted proteins expressed in haustoria (Petre *et al*., 2015)	24	Small size	–	–	–
		Small, cysteine-rich	15 (62.5%)	707 (38.8%)	**0.0210**
		EffectorP 1.0	20 (83.3%)	780 (42.8%)	<**0.0001**
		EffectorP 2.0	18 (75%)	752 (41.3%)	**0.0013**
Melampsora larici-populina: specific small secreted proteins expressed during infection (Germain et al., 2011)	16	Small size	–	–	–
		Small, cysteine-rich	10 (62.5%)	707 (38.8%)	NS
		EffectorP 1.0	15 (93.8%)	780 (42.8%)	<**0.0001**
		EffectorP 2.0	14 (87.5%)	752 (41.3%)	**0.0004**
Laccaria bicolor: ectomycorrhiza-regulated small secreted proteins (MiSSPs) (Martin *et al*., 2008)	21	Small size	–	–	–
		Small, cysteine-rich	10 (47.6%)	362 (29.2%)	NS
		EffectorP 1.0	11 (52.4%)	380 (30.7%)	NS
		EffectorP 2.0	10 (47.6%)	246 (19.9%)	**0.0043**
Puccinia graminis f. sp. *tritici*: secreted proteins up-regulated in haustoria (log FC $>$ 10) (Upadhyaya *et al*., 2015)	55	Small size	38 (69.1%)	1223 (64.7%)	NS
		Small, cysteine-rich	7 (12.7%)	710 (37.6%)	NS
		EffectorP 1.0	25 (45.5%)	841 (44.5%)	NS
		EffectorP 2.0	22 (40%)	758 (40.1%)	NS
Zymoseptoria tritici candidate effectors (Kettles et al., 2017)	63	Small size	56 (88.9%)	426 (42.7%)	<**0.0001**
		Small, cysteine-rich	43 (68.3%)	259 (26%)	<**0.0001**
		EffectorP 1.0	41 (65.1%)	260 (26.1%)	<**0.0001**
		EffectorP 2.0	42 (66.7%)	232 (23.3%)	<**0.0001**
Zymoseptoria tritici candidate effectors with phenotype in *Nicotiana benthamiana* (Kettles et al., 2017)	14	Small size	12 (85.7%)	426 (42.7%)	**0.0017**
		Small, cysteine-rich	10 (71.4%)	259 (26%)	**0.0005**
		EffectorP 1.0	8 (57.1%)	260 (26.1%)	**0.0143**
		EffectorP 2.0	9 (64.3%)	232 (23.3%)	**0.0014**
Ustilago maydis effector candidates (Tollot *et al*., 2016)	198	Small size	130 (65.7%)	242 (46.8%)	<**0.0001**
		Small, cysteine-rich	49 (24.7%)	101 (19.5%)	NS
		EffectorP 1.0	79 (39.9%)	140 (27.1%)	**0.0011**
		EffectorP 2.0	67 (33.8%)	124 (24%)	**0.0106**
Leptosphaeria maculans highly expressed early effector candidates (Gervais *et al*., 2017)	49	Small size	44 (89.9%)	514 (49.7%)	<**0.0001**
		Small, cysteine-rich	23 (46.9%)	258 (24.9%)	**0.0013**
		EffectorP 1.0	29 (59.2%)	283 (27.3%)	<**0.0001**
		EffectorP 2.0	26 (53.1%)	215 (20.8%)	<**0.0001**
Leptosphaeria maculans highly expressed late effector candidates (Gervais *et al*., 2017)	50	Small size	33 (66%)	514 (49.7%)	**0.0292**
		Small, cysteine-rich	16 (32%)	258 (24.9%)	NS
		EffectorP 1.0	19 (38%)	283 (27.3%)	NS
		EffectorP 2.0	19 (38%)	215 (20.8%)	**0.0073**

For each expression dataset, the percentage of predicted effector candidates by EffectorP is shown and compared with the percentage of predicted effector candidates in the secretome. The small size classifier is only applied to sets that are not pre-selected based on a small size.

Table 6 Predicted effectors in secretomes for groups of fungal species

	Average of predicted effectors		
Secretomes	EffectorP 1.0	EffectorP 2.0	% decrease in predicted effectors (EffectorP 2.0 compared with EffectorP 1.0)
Plant pathogens	338 (29.6%)	284 (24.9%)	−16.0%
Fungal symbionts of plants	305 (30.8%)	177 (17.8%)	−42.0%
Fungal pathogens of animals	108 (20.9%)	83 (16.1%)	−23.2%
Saprophytes	177 (19.5%)	106 (11.7%)	−40.1%

(37.6%). Amongst the fungal plant pathogens, the lowest proportions of predicted effectors were recorded for the necrotrophs *Heterobasidion annosum* (10.4%), *Sclerotinia sclerotiorum* (13.6%), *Botrytis cinerea* (13.7%) and *Penicillium digitatum* (13.9%). Necrotrophic pathogens utilize many secreted PCWDEs to overcome the barrier of the plant cell wall. EffectorP predicts some secreted proteins with enzymatic domains as effectors, such as the *Fusarium graminearum* xylanase XylA, which has the ability to induce necrosis in wheat independent of its enzymatic activity (Table 4) (Belien *et al.*, 2011; Sella *et al.*, 2016; Sperschneider *et al.*, 1996). However, EffectorP has been trained on effectors that predominantly lack recognizable functional domains and interfere with host processes in different ways from PCWDEs which act on the plant cell wall. Therefore, the lower proportions of EffectorP-predicted effectors in necrotrophic fungal pathogen secretomes is expected.

On average, EffectorP 2.0 predicts that plant pathogen secretomes consist of 24.9% effectors and that saprophyte secretomes consist of 11.7% effectors (Tables 5 and 6). EffectorP 2.0 reduces the average number of predicted effectors in fungal plant symbiont and fungal saprophyte secretomes by over 40% when compared with EffectorP 1.0 (Table 6, Fig. 3). Both EffectorP 2.0 and EffectorP 1.0 also predict lower proportions of effectors for fungal animal pathogens than for fungal plant pathogens (Table 6),

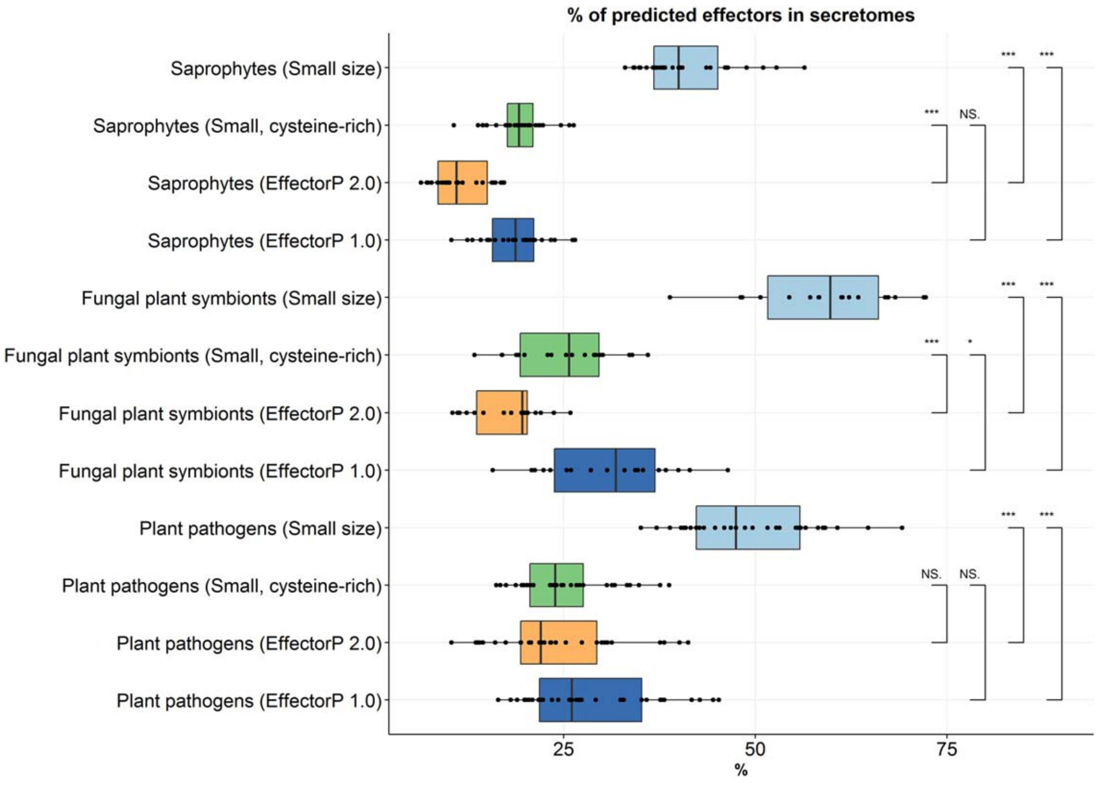

Fig. 3 Proportions of predicted effectors in fungal secretomes using EffectorP 1.0, EffectorP 2.0, a small size classifier and a small, cysteine-rich classifier. All data points are drawn on top of the box plots as black dots. Significance between groups is shown as horizontal brackets and was assessed using *t*-tests (NS, not significant; $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$). The lower and upper hinges correspond to the first and third quartiles and the upper (lower) whiskers extend from the hinge to the largest (smallest) value that is within 1.5 times the interquartile range of the hinge. Data beyond the end of the whiskers are outliers.

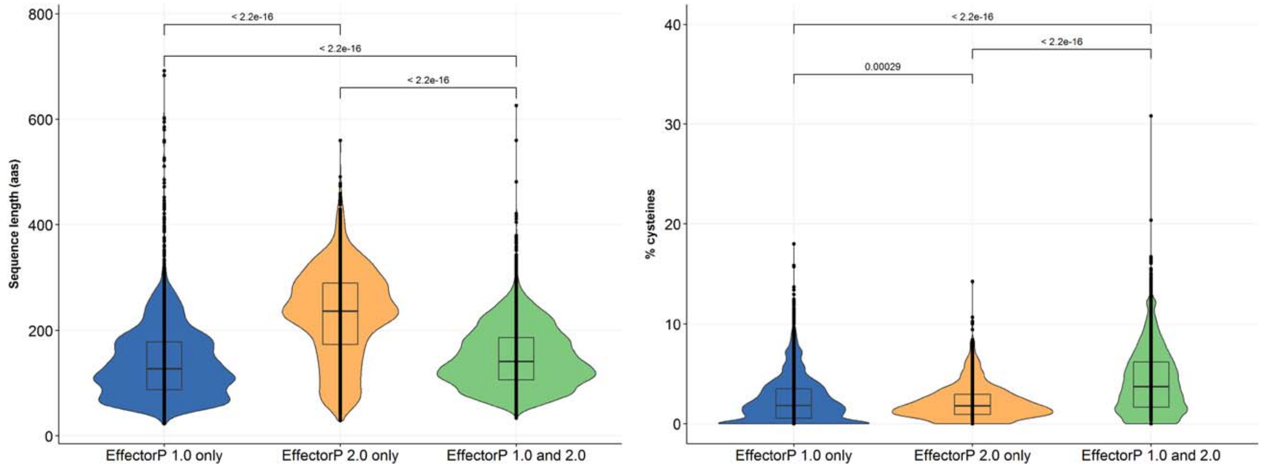

Fig. 4 Differences in sequence length (aas, amino acids) and cysteine content for effectors predicted by different versions of EffectorP. All data points are drawn on top of the box plots as black dots. Significance between groups is shown as horizontal brackets and was assessed using *t*-tests. The lower and upper hinges correspond to the first and third quartiles and the upper (lower) whiskers extend from the hinge to the largest (smallest) value that is within 1.5 times the interquartile range of the hinge. Data beyond the end of the whiskers are outliers.

suggesting that effector repertoires of fungal animal pathogens are different from those of their plant-pathogenic counterparts. One notable exception is the secretome of *Enterocytozoon bieneusi*, an obligate intracellular parasite (49 predicted effectors, 36% of secretome predicted as effectors). Shortened protein-coding sequences caused by genome compaction have been reported in *E. bieneusi* (Akiyoshi *et al*., 2009) and might lead to higher than expected false positive predictions. Therefore, we also assessed effector prediction rates for small secreted proteins (<300 amino acids) only. For plant pathogens, EffectorP 2.0 predicts that 47.8% of small secreted proteins are effectors, whereas, for plant symbionts and saprophytes, this is reduced to 29.9% and 26.3%, respectively. This underlines that EffectorP 2.0 does not select effectors based on a small size alone. Small secreted proteins in saprophytes are mostly functionally uncharacterized and might function in a variety of processes unrelated to plant–pathogen interactions. Compared with a small, cysteine-rich classifier, EffectorP 2.0 predicts significantly lower proportions of effectors for plant symbionts and saprophytes, but not for plant pathogens (Fig. 3). This lack of correlation for all groups tested underlines that EffectorP 2.0 does not select effectors based on a small size and a high cysteine content alone, and reflects the reduced false positive rate of EffectorP 2.0.

We then further investigated the properties of effectors that are only predicted by one of the versions of EffectorP, but not by the other, for all 93 secretomes (Table S4). Effector candidates predicted only by EffectorP 2.0 are, on average, of longer sequence length ($n = 2304$; average sequence length, 229 amino acids) than those that are only predicted by EffectorP 1.0 ($n = 8635$; average sequence length, 138 amino acids) or by both versions ($n = 14\,128$; average sequence length, 148 amino acids) (Fig. 4). Effector candidates predicted only by EffectorP 1.0 or 2.0 are lower in cysteine content compared with effector candidates predicted by both versions (Fig. 4). We then tested for enrichment and depletion of protein functional classes amongst the effector candidates predicted by EffectorP 1.0 and 2.0 from a total of 24 075 secreted proteins of the 21 plant pathogens (Table 7). The vast majority of effector candidates predicted by either EffectorP 1.0 or 2.0 are proteins without functional annotation. However, we observed that both sets of predicted effector candidates are enriched for proteins with pectate lyase activity, peptidyl-prolyl *cis–trans* isomerase activity and endopeptidase inhibitor activity (Table 7). Some proteins with peptidyl-prolyl *cis–trans* isomerase activity have been implicated to function as virulence factors (Unal and Steinert, 2009). A cyclophilin with peptidyl-prolyl *cis–trans* isomerase activity functions as a pathogenicity factor in *Puccinia triticina* (Panwar *et al*., 2012). EffectorP 2.0-predicted effectors are enriched for proteins involved in pathogenesis and defence response (Table 7). However, EffectorP 1.0-predicted effector candidates are also enriched for proteins that do not appear to be related to effector function or to secreted proteins, but rather to intracellular proteins (Table 7), and might reflect the higher false positive rate of EffectorP 1.0, as well as the false positive rate of the signal peptide prediction tools SignalP 3.0 and TargetP.

DISCUSSION

Given the high diversity of fungal effectors, it seems an unexpected finding that a machine learning classifier can accurately distinguish diverse effectors from secreted non-effectors. However, classifiers such as decision trees can have multiple paths that lead to a prediction as an effector and one can speculate that

Table 7 Gene ontology (GO) term enrichment analysis of predicted effector candidates.

Comparison	Over-represented GO term description	# proteins in test set	# proteins in reference set	FDR
Test set: EffectorP 2.0 predicted	Pectate lyase activity	26	11	2.9×10^{-8}
Reference set: Secreted pathogen proteins	Peptidyl-prolyl *cis–trans* isomerase activity	15	5	2.6×10^{-5}
	Pathogenesis	12	10	0.02
	Defence response	14	14	0.02
	Endopeptidase inhibitor activity	7	3	0.03
Test set: EffectorP 1.0 predicted	Peptidyl-prolyl *cis–trans* isomerase activity	18	2	1.9×10^{-7}
Reference set: Secreted pathogen proteins	Inner mitochondrial membrane organization	10	1	3.6×10^{-4}
	Intracellular sterol transport	8	0	5.7×10^{-4}
	Fungal-type vacuole lumen	8	0	5.7×10^{-4}
	Pectate lyase activity	22	15	8.4×10^{-4}
	Endopeptidase inhibitor activity	9	1	0.001
	Chaperone-mediated protein folding	7	0	0.001
	FK506 binding	7	0	0.002
	Nuclear envelope organization	6	0	0.002
	Regulation of COPII vesicle coating	6	0	0.006
	Endoplasmic reticulum exit site	6	0	0.006
	Mitochondrial inner membrane	18	15	0.01
	Mitochondrial respiratory chain complex IV assembly	5	0	0.02
	Mitochondrion morphogenesis	5	0	0.02
	COPII vesicle coat	6	1	0.03

different paths might relate to different classes of effectors, such as apoplastic or cytoplasmic effectors. Decision trees can also learn feature interactions, whereas Naïve Bayes classifiers identify the importance of individual features, but not relationships amongst features. This might be advantageous for effector prediction, e.g. a Naïve Bayes classifier can learn that a small protein size or a high cysteine content is important for effectors, but it does not learn that proteins have to be small and at the same time cysteine-rich to be effectors. Unlike Naïve Bayes classifiers, decision trees are non-parametric, which gives them the ability to,

Table 8 Genomes that were used to predict secretomes for negative training data.

Ecology	Species	Reference
Fungal pathogen/symbiont	*Melampsora lini*, *Puccinia graminis* f. sp. *tritici*, *P. striiformis* f. sp. *tritici*, *Blumeria graminis* f. sp. *hordei*, *B. graminis* f. sp. *tritici*, *Cladosporium fulvum*, *Leptosphaeria maculans*, *Fusarium oxysporum* f. sp. *lycopersici*, *F. graminearum*, *Magnaporthe oryzae*, *Rhynchosporium secalis*, *Verticillium dahliae*, *Ustilago maydis*, *U. hordei*, *Stagonospora nodorum*, *Botrytis cinerea*, *Pyrenophora tritici-repentis*, *Laccaria bicolor*, *Zymoseptoria tritici*, *Colletotrichum graminicola*, *Sclerotinia sclerotiorum*	Nemri *et al.* (2014), Duplessis *et al.* (2011), Cantu *et al.* (2011), Spanu *et al.* (2010), Wicker *et al.* (2013), de Wit *et al.* (2012), Rouxel *et al.* (2011), Ma *et al.* (2010), Cuomo *et al.* (2007), Dean *et al.* (2005), Penselin *et al.* (2016), Klosterman *et al.* (2011), Kämper *et al.* (2006), Laurie *et al.* (2012), Hane *et al.* (2007), Amselem *et al.* (2011), Manning *et al.* (2013), Martin *et al.* (2008), Goodwin *et al.* (2011), O'Connell *et al.* (2012)
Fungal saprophyte	*Agaricus bisporus* var. *bisporus*, *Amanita thiersii*, *Aspergillus niger*, *A. oryzae*, *Coniophora puteana*, *Dacryopinax* sp., *Dichomitus squalens*, *Fomitiporia mediterranea*, *Fomitopsis pinicola*, *Gloeophyllum trabeum*, *Punctularia strigosozonata*, *Stereum hirsutum*, *Trametes versicolor*, *Wolfiporia cocos*, *Gymnopus luxurians*, *Hydnomerulius pinastri*, *Hypholoma sublateritium*, *Plicaturopsis crispa*, *Sphaerobolus stellatus*, *Hysterium pulicare*, *Neurospora crassa*, *Pichia stipitis*, *Pseudozyma antarctica*, *P. aphidis*, *Rhodosporidium toruloides*, *Saccharomyces cerevisiae*, *Coprinus cinereus*	Morin *et al.* (2012), Hess *et al.* (2014), Andersen *et al.* (2011), Machida *et al.* (2005), Floudas *et al.* (2012), Kohler *et al.* (2015), Ohm *et al.* (2014), Galagan *et al.* (2003), Jeffries *et al.* (2007), Morita *et al.* (2013), Lorenz *et al.* (2014), Zhu *et al.* (2012), Goffeau *et al.* (1996), Stajich *et al.* (2010)
Animal pathogen	*Batrachochytrium dendrobatidis*, *Candida albicans*, *Cordyceps militaris*, *Cryptococcus neoformans* var. *grubii*, *C. neoformans* var. *neoformans*, *Encephalitozoon cuniculi*, *Enterocytozoon bieneusi*, *Malassezia globosa*, *Metarhizium robertsii*, *Paracoccidioides brasiliensis*	Rosenblum *et al.* (2010), Jones *et al.* (2004), Zheng *et al.* (2011), Loftus *et al.* (2005), Janbon *et al.* (2014), Katinka *et al.* (2001), Akiyoshi et al. (2009), Xu *et al.* (2007), Gao *et al.* (2011), Desjardins *et al.* (2011)

for example, assign a very low protein size to non-effectors, a low to medium protein size to effectors and a large protein size to non-effectors. However, decision trees are prone to overfitting, especially on small training datasets, which can lead to a limited ability to correctly classify unseen data. Naïve Bayes classifiers can deliver robust performance on small training datasets and an ensemble classifier, such as EffectorP 2.0, is capable of drawing on the strengths of both decision trees and Naïve Bayes classifiers.

On the current training set, low molecular weight is an important feature in fungal effector classification. However, it is possible that fungal pathogens employ classes of larger effector proteins which have thus far not been recognized. For example, the recently discovered *Puccinia graminis* f. sp. *tritici* effectors AvrSr50 (Chen *et al.*, 2003) and AvrSr35 (Salcedo *et al.*, 2005) are 132 and 578 amino acids long, respectively. With sufficient training data, EffectorP could learn to recognize classes of effectors that share no sequence similarity, yet are structurally conserved, such as MAX-effectors (de Guillen *et al.*, 2016). Machine learning classifiers trained to recognize oomycete RxLR effectors could be used to search for effectors with similar structural properties in fungi. In general, future re-training of EffectorP on the expanding sets of experimentally supported effectors will be critical to retain its value. We envisage that, in the future, separate training sets of apoplastic fungal effectors and cytoplasmic fungal effectors could be of sufficient size to allow for the training of separate classifiers, which could potentially increase prediction accuracy. Although the machine learning classifier ApoplastP delivers accurate prediction of apoplastic protein localization for both plant and effector proteins (Sperschneider *et al.*, 2008), other signals unique to apoplastic or cytoplasmic effectors might not be fully utilized by EffectorP as yet.

Another challenge is the choice of the negative training set, which should ideally contain no undiscovered effectors. However, the set of secreted fungal pathogen proteins is mostly unlabelled and will contain true positive effectors. To minimize this effect, we filtered the predicted pathogen secretomes for EffectorP 1.0-predicted effector candidates, which removed predominantly small, cysteine-rich effector candidates. This could introduce the possibility that a classifier trained on fungal effectors (many are small, cysteine-rich proteins) and EffectorP 1.0-filtered secreted pathogen proteins (many small, cysteine-rich proteins removed) would bias a classifier towards the recognition of predominantly small, cysteine-rich proteins as effectors. However, this does not seem to be the case for EffectorP 2.0. Although machine learning classifiers can, to some degree, be tolerant to noisy negative training data, in particular if the positive set is of high quality, undiscovered effectors might remain in the negative set and potentially bias predictions.

Practical recommendations for fungal effector prediction depend on the application. For example, for subsequent experimental validation in which time and resources are limited, a stringent effector screening approach might be most appropriate. This could involve taking either EffectorP 2.0-predicted effectors, or effectors predicted by both versions of EffectorP 1.0/2.0 for maximum stringency. For maximum sensitivity, a union of effector candidates predicted by either EffectorP 1.0 or 2.0 could be used; however, this will also result in high false positive rates. If *in planta* expression data are available, effectors expressed highly during infection can be prioritized for experimental validation. Another approach would be to select effectors with highest probability; however, this has not been tested extensively by us. Nevertheless, we did observe that, during the identification of the *Puccinia graminis* f. sp. *tritici* effector AvrSr50 (Chen *et al.*, 2003), where over 40 candidate genes had to be functionally screened, the application of EffectorP 2.0 and ApoplastP (Sperschneider *et al.*, 2008) to predict the most likely effector to enter plant cells would have revealed AvrSr50 as the top candidate with highest probability. Overall, the re-evaluation and re-training of EffectorP have supported the power of machine learning for fungal effector prediction. Higher accuracy of fungal effector prediction will boost experimental validation success rates and aid in the understanding of effector biology.

EXPERIMENTAL PROCEDURES

Training of the machine learning classifier

As a positive training set, we collected validated fungal effectors from the literature and then reduced sequence homology in this set by removing those that shared similarity with another effector in the set at bit score ≥ 50 using phmmer (Finn *et al.*, 2015). Three negative training sets were generated based on secretomes predicted from annotated gene sets of publicly available genome assemblies of plant-pathogenic fungi and symbionts (21 species, same species from the positive effector training set), animal-pathogenic fungi (10 species) or saprophytic fungi (27 species) (Table 8). A protein was labelled as secreted if it was predicted to be secreted by the neural network predictor of SignalP 3 (Bendtsen *et al.*, 2011) as well as by TargetP (Emanuelsson *et al.*, 2006), and if it had no predicted transmembrane domain outside the first 60 amino acids using TMHMM (Krogh *et al.*, 2015), as described previously for fungal effector prediction (Sperschneider *et al.*, 2011). Each negative set was homology reduced by deleting proteins that shared sequence similarity (bit score ≤ 100, phmmer) with another in the negative set. We also applied EffectorP 1.0 (Sperschneider *et al.*, 1996) to exclude predicted effectors from the fungal pathogen/symbiont secretomes. The WEKA tool box (version 3.8.1) was used to train and evaluate the performance of different machine learning classifiers (Hall *et al.*, 2000), and feature vectors were calculated for each protein (Table 2). The training data are available at: http://effectorp.csiro.au/data.html.

For the ensemble learner, we took 100 randomly selected samples of negative training data from each of the three negative sets (pathogen/symbiont secretomes, saprophyte secretomes and animal pathogen secretomes), each with 282 protein sequences, to give a ratio of 3 : 1 to the number of positive training examples. We then used WEKA to train Naïve

Bayes classifiers on each of the 300 negative datasets with the same positive training set. We then repeated this procedure and trained C4.5 decision trees (J48 model in WEKA) on another 300 randomly chosen negative datasets from the three classes. For each set of 100 models, we selected the best-performing models as those with the highest area under the curve (AUC). Overall, we chose a total of 50 models comprising: 10 Naïve Bayes classifiers and 10 C4.5 decision trees that discriminate between fungal effectors and secreted pathogen proteins; 10 Naïve Bayes classifiers and 10 C4.5 decision trees that discriminate between fungal effectors and secreted saprophyte proteins; and five Naïve Bayes classifiers and five C4.5 decision trees that discriminate between fungal effectors and secreted animal pathogen proteins. The ensemble classifier called EffectorP 2.0 returns a final prediction using a soft voting approach, which predicts the class label based on average probabilities for 'effector' and 'non-effector' calculated by each classifier. Soft voting then returns the class with the highest average probability as the result. A protein is classified as an effector if it has a probability > 0.55. If it is predicted as an effector with probability 0.5–0.55, it is labelled as an 'unlikely effector' and is counted as a non-effector in the evaluation.

Evaluation of EffectorP 2.0

We collected fungal, plant and mammalian proteins with experimentally validated localization to endoplasmic reticulum, Golgi or membranes or with GPI anchors from the UniProt database (search terms in Table S1, see Supporting Information), and predicted signal peptides using SignalP 4.1 (Petersen *et al.*, 2003). We also collected fungal proteins from PHI-base (Urban *et al.*, 2007) from *Fusarium*, *Magnaporthe*, *Ustilago*, *Sclerotinia*, *Botrytis*, *Zymoseptoria* and *Leptosphaeria* pathogens, which are annotated as having an unaffected pathogenicity phenotype. All evaluation data are available at: http://effectorp.csiro.au/data.html.

In the evaluation, true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs) are calculated. Accuracy is reported as (TP + TN)/(TP + TN + FP + FN), whereas sensitivity is the fraction of effectors that are correctly identified as such [TP/(TP + FN)] and specificity is the fraction of non-effectors which are correctly identified as such [TN/(TN + FP)]. The positive predictive value (PPV) is the proportion of positive results that are true positives [TP/(TP + FP)]. Receiver operating characteristic (ROC) curves plot sensitivity against (1 – specificity) and the area under the curve (AUC) can be calculated. This value gives the probability that a classifier will rank a randomly chosen effector higher than a randomly chosen non-effector. Therefore, a perfect classifier achieves an AUC of 1.0, whereas a random classifier achieves an AUC of only 0.5.

A small size classifier predicts a protein as an effector if it has a sequence length of $\leq$300 amino acids, and a small, cysteine-rich classifier predicts a protein as an effector if it has a sequence length of $\leq$300 amino acids and $\geq$4 cysteines in its sequence.

Functional enrichment analysis and plotting

We performed sequence similarity searches against fungal proteins at the National Center for Biotechnology Information (NCBI) with Blast2GO 4.1.9 (Gotz *et al.*, 2011) and default parameters. GO terms were reduced to the most specific terms and Fisher's exact tests were used to find over- and under-represented terms. Enrichment was called at false discovery rate (FDR) < 0.05.

Plots were produced using ggplot2 (Wickham, 2009) and statistical significance was assessed with *t*-tests using the ggsignif package (https://cran.r-project.org/web/packages/ggsignif/index.html). Significance thresholds according to *t*-test are NS = not significant, $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$.

ACKNOWLEDGEMENTS

J.S. was supported by a CSIRO Office of the Chief Executive (OCE) Post-doctoral Fellowship. We thank Jonathan Anderson and Jonathan Powell for comments on the manuscript.

AUTHOR CONTRIBUTIONS

J.S. planned and designed the research and developed the software. All authors analysed the data and wrote the manuscript.

REFERENCES

Akiyoshi, D.E., Morrison, H.G., Lei, S., Feng, X., Zhang, Q., Corradi, N., Mayanja, H., Tumwine, J.K., Keeling, P.J., Weiss, L.M. and Tzipori, S. (2009) Genomic survey of the non-cultivatable opportunistic human pathogen, Enterocytozoon bieneusi. *PLoS Pathog.* **5**, e1000261.

Amselem, J., Cuomo, C.A., van Kan, J.A.L., Viaud, M., Benito, E.P., Couloux, A., Coutinho, P.M., de Vries, R.P., Dyer, P.S., Fillinger, S., Fournier, E., Gout, L., Hahn, M., Kohn, L., Lapalu, N., Plummer, K.M., Pradier, J.-M., Quévillon, E., Sharon, A., Simon, A., ten Have, A., Tudzynski, B., Tudzynski, P., Wincker, P., Andrew, M., Anthouard, V., Beever, R.E., Beffa, R., Benoit, I., Bouzid, O., Brault, B., Chen, Z., Choquer, M., Collémare, J., Cotton, P., Danchin, E.G., Da Silva, C., Gautier, A., Giraud, C., Giraud, T., Gonzalez, C., Grossetete, S., Güldener, U., Henrissat, B., Howlett, B.J., Kodira, C., Kretschmer, M., Lappartient, A., Leroch, M., Levis, C., Mauceli, E., Neuvéglise, C., Oeser, B., Pearson, M., Poulain, J., Poussereau, N., Quesneville, H., Rascle, C., Schumacher, J., Ségurens, B., Sexton, A., Silva, E., Sirven, C., Soanes, D.M., Talbot, N.J., Templeton, M., Yandava, C., Yarden, O., Zeng, Q., Rollins, J.A., Lebrun, M.-H. and Dickman, M. (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genet.* **7**, e1002230.

Andersen, M.R., Salazar, M.P., Schaap, P.J., van de Vondervoort, P.J.I., Culley, D., Thykaer, J., Frisvad, J.C., Nielsen, K.F., Albang, R., Albermann, K., Berka, R.M., Braus, G.H., Braus-Stromeyer, S.A., Corrochano, L.M., Dai, Z., van Dijck, P.W.M., Hofmann, G., Lasure, L.L., Magnuson, J.K., Menke, H., Meijer, M., Meijer, S.L., Nielsen, J.B., Nielsen, M.L., van Ooyen, A.J.J., Pel, H.J., Poulsen, L., Samson, R.A., Stam, H., Tsang, A., van den Brink, J.M., Atkins, A., Aerts, A., Shapiro, H., Pangilinan, J., Salamov, A., Lou, Y., Lindquist, E., Lucas, S., Grimwood, J., Grigoriev, I.V., Kubicek, C.P., Martinez, D., van Peij, N.N.M.E., Roubos, J.A., Nielsen, J. and Baker, S.E. (2011) Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res.* **21**, 885–897.

Belien, T., Van Campenhout, S., Van Acker, M., Robben, J., Courtin, C.M., Delcour, J.A. and Volckaert, G. (2007) Mutational analysis of endoxylanases XylA and XylB from the phytopathogen *Fusarium graminearum* reveals comprehensive insights into their inhibitor insensitivity. *Appl. Environ. Microbiol.* **73**, 4602–4608.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795.

Bing, L., Yang, D., Li, X.L., Lee, W.S. and Yu, P.S. (2003) Building text classifiers using positive and unlabeled examples. In: *Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, USA*, pp. 179–186.

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K.K., Jurka, J., Michelmore, R.W. and Dubcovsky, J. (2011) Next generation sequencing provides rapid access to the genome of Puccinia striiformis f. sp. tritici, the causal agent of wheat stripe rust. *PLoS One.* **6**, e24230.

Catanzariti, A.M., Dodds, P.N., Lawrence, G.J., Ayliffe, M.A. and Ellis, J.G. (2006) Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell*, **18**, 243–256.

Chen, J., Upadhyaya, N.M., Ortiz, D., Sperschneider, J., Li, F., Bouton, C., Breen, S., Dong, C., Xu, B., Zhang, X., Mago, R., Newell, K., Xia, X., Bernoux, M., Taylor, J.M., Steffenson, B., Jin, Y., Zhang, P., Kanyuka, K., Figueroa, M., Ellis, J.G., Park, R.F. and Dodds, P.N. (2017) Loss of AvrSr50 by somatic exchange in stem rust leads to virulence for Sr50 resistance in wheat. *Science*, **358**, 1607–1610.

Cuomo, C.A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B.G., Di Pietro, A., Walton, J.D., Ma, L.-J., Baker, S.E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y.-L., Decaprio, D., Gale, L.R., Gnerre, S., Goswami, R.S., Hammond-Kosack, K., Harris, L.J., Hilburn, K., Kennell, J.C., Kroken, S., Magnuson, J.K., Mannhaupt, G., Mauceli, E., Mewes, H.-W., Mitterbauer, R., Muehlbauer, G., Münsterkötter, M., Nelson, D., O'donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M.I.G., Seong, K.-Y., Tetko, I.V., Urban, M., Waalwijk, C., Ward, T.J., Yao, J., Birren, B.W. and Kistler, H.C. (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.-R., Pan, H., Read, N.D., Lee, Y.-H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.-H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.-J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E. and Birren, B.W. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.

Desjardins, C.A., Champion, M.D., Holder, J.W., Muszewska, A., Goldberg, J., Bailão, A.M., Brigido, M.M., Ferreira, M. E D S., Garcia, A.M., Grynberg, M., Gujja, S., Heiman, D.I., Henn, M.R., Kodira, C.D., León-Narváez, H., Longo, L.V.G., Ma, L.-J., Malavazi, I., Matsuo, A.L., Morais, F.V., Pereira, M., Rodríguez-Brito, S., Sakthikumar, S., Salem-Izacc, S.M., Sykes, S.M., Teixeira, M.M., Vallejo, M.C., Walter, M.E.M.T., Yandava, C., Young, S., Zeng, Q., Zucker, J., Felipe, M.S., Goldman, G.H., Haas, B.J., McEwen, J.G., Nino-Vega, G., Puccia, R., San-Blas, G., Soares, C. M D A., Birren, B.W. and Cuomo, C.A. (2011) Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. *PLoS Genet.* **7**, e1002345.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, CHee., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C. and Obradovic, Z. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59.

Duplessis, S., Cuomo, C.A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D.L., Hacquard, S., Amselem, J., Cantarel, B.L., Chiu, R., Coutinho, P.M., Feau, N., Field, M., Frey, P., Gelhaye, E., Goldberg, J., Grabherr, M.G., Kodira, C.D., Kohler, A., Kües, U., Lindquist, E.A., Lucas, S.M., Mago, R., Mauceli, E., Morin, E., Murat, C., Pangilinan, J.L., Park, R., Pearson, M., Quesneville, H., Rouhier, N., Sakthikumar, S., Salamov, A.A., Schmutz, J., Selles, B., Shapiro, H., Tanguay, P., Tuskan, G.A., Henrissat, B., Van de Peer, Y., Rouzé, P., Ellis, J.G., Dodds, P.N., Schein, J.E., Zhong, S., Hamelin, R.C., Grigoriev, I.V., Szabo, L.J. and Martin, F. (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U S A.* **108**, 9166–9171.

Ellis, J.G. (2016) Integrated decoys and effector traps: how to catch a plant pathogen. *BMC Biol.* **14**, 13.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016.

Fauchere, J.-L. and Pliska, V.E. (1983) Hydrophobic parameters pi of amino acid side chains from partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.

Figueroa, M., Hammond-Kosack, K.E. and Solomon, P.S. (2017) A review of wheat diseases - a field perspective. *Mol. Plant. Pathol.* doi: 10.1111/mpp.12618.

Figueroa, M., Manning, V.A., Pandelova, I. and Ciuffetti, L.M. (2015) Persistence of the Host-Selective Toxin Ptr ToxB in the Apoplast. *Mol. Plant-Microbe Interact.* **28**, 1082–1090.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37.

Fischer, R.A., Byerlee, D. and Edmeades, G.O. 2014. *Crop Yields and Global Food Security: will Yield Increase Continue to Feed the World*? Canberra: Australian Centre for International Agricultural Research.

Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R.A., Henrissat, B., Martínez, A.T., Otillar, R., Spatafora, J.W., Yadav, J.S., Aerts, A., Benoit, I., Boyd, A., Carlson, A., Copeland, A., Coutinho, P.M., de Vries, R.P., Ferreira, P., Findley, K., Foster, B., Gaskell, J., Glotzer, D., Górecki, P., Heitman, J., Hesse, C., Hori, C., Igarashi, K., Jurgens, J.A., Kallen, N., Kersten, P., Kohler, A., Kües, U., Kumar, T.K.A., Kuo, A., LaButti, K., Larrondo, L.F., Lindquist, E., Ling, A., Lombard, V., Lucas, S., Lundell, T., Martin, R., McLaughlin, D.J., Morgenstern, I., Morin, E., Murat, C., Nagy, L.G., Nolan, M., Ohm, R.A., Patyshakuliyeva, A., Rokas, A., Ruiz-Dueñas, F.J., Sabat, G., Salamov, A., Samejima, M., Schmutz, J., Slot, J.C., St John, F., Stenlid, J., Sun, H., Sun, S., Syed, K., Tsang, A., Wiebenga, A., Young, D., Pisabarro, A., Eastwood, D.C., Martin, F., Cullen, D., Grigoriev, I.V. and Hibbett, D.S. (2012) The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science*, **336**, 1715–1719.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R.L., Perkins, D.D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P.C., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S., Lander, E.S., Nusbaum, C. and Birren, B. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.

Gao, Q., Jin, K., Ying, S.-H., Zhang, Y., Xiao, G., Shang, Y., Duan, Z., Hu, X., Xie, X.-Q., Zhou, G., Peng, G., Luo, Z., Huang, W., Wang, B., Fang, W., Wang, S., Zhong, Y., Ma, L.-J., St. Leger, R.J., Zhao, G.-P., Pei, Y., Feng, M.-G., Xia, Y. and Wang, C. (2011) Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi Metarhizium anisopliae and M. acridum. *PLoS Genet.* **7**, e1001264.

Garnica, D.P., Nemri, A., Upadhyaya, N.M., Rathjen, J.P. and Dodds, P.N. (2014) The ins and outs of rust haustoria. *PLoS Pathog.* **10**, e1004329.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM ed. pp. 571–607. *The Proteomics Protocols Handbook*: Humana Press.

Germain, H., Joly, D.L., Mireault, C., Plourde, M.B., Letanneur, C., Stewart, D., Morency, M.J., Petre, B., Duplessis, S. and Seguin, A. (2016) Infection assays in Arabidopsis reveal candidate effectors from the poplar rust fungus that promote susceptibility to bacteria and oomycete pathogens. *Mol. Plant. Pathol.*

Gervais, J., Plissonneau, C., Linglin, J., Meyer, M., Labadie, K., Cruaud, C., Fudal, I., Rouxel, T. and Balesdent, M.H. (2017) Different waves of effector genes with contrasted genomic location are expressed by Leptosphaeria maculans during cotyledon and stem colonization of oilseed rape. *Mol. Plant. Pathol.* **18**, 1113–1126.

Gibson, D.M., King, B.C., Hayes, M.L. and Bergstrom, G.C. (2011) Plant pathogens as a source of diverse enzymes for lignocellulose digestion. *Curr Opin Microbiol.* **14**, 264–270.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes. *Science.* **274**, 546–563. 547.

Goodwin, S.B., Ben M'Barek, S., Dhillon, B., Wittenberg, A.H.J., Crane, C.F., Hane, J.K., Foster, A.J., Van der Lee, T.A.J., Grimwood, J., Aerts, A., Antoniw, J., Bailey, A., Bluhm, B., Bowler, J., Bristow, J., van der Burgt, A., Canto-Canché, B., Churchill, A.C.L., Conde-Ferràez, L., Cools, H.J., Coutinho, P.M., Csukai, M., Dehal, P., De Wit, P., Donzelli, B., van de Geest, H.C., van Ham, R.C.H.J., Hammond-Kosack, K.E., Henrissat, B., Kilian, A., Kobayashi, A.K., Koopmann, E., Kourmpetis, Y., Kuzniar, A., Lindquist, E., Lombard, V., Maliepaard, C., Martins, N., Mehrabi, R., Nap, J.P.H., Ponomarenko, A., Rudd, J.J., Salamov, A., Schmutz, J., Schouten, H.J., Shapiro, H., Stergiopoulos, I., Torriani, S.F.F., Tu, H., de Vries, R.P., Waalwijk, C., Ware, S.B., Wiebenga, A., Zwiers, L.-H., Oliver, R.P., Grigoriev, I.V. and Kema, G.H.J. (2011) Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* **7**, e1002070.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J. and Conesa, A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435.

de Guillen, K., Ortiz-Vallejo, D., Gracy, J., Fournier, E., Kroj, T. and Padilla, A. (2015) Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLoS Pathog.* **11**, e1005228.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newslett* **11**, 10–18.

Hane, J.K., Lowe, R.G.T., Solomon, P.S., Tan, K.-C., Schoch, C.L., Spatafora, J.W., Crous, P.W., Kodira, C., Birren, B.W., Galagan, J.E., Torriani, S.F.F., McDonald, B.A. and Oliver, R.P. (2007) *Dothideomycete* plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum. Plant Cell.* **19**, 3347–3368.

Hess, J., Skrede, I., Wolfe, B.E., LaButti, K., Ohm, R.A., Grigoriev, I.V. and Pringle, A. (2014) Transposable element dynamics among asymbiotic and ectomycorrhizal Amanita fungi. *Genome Biol Evol.* **6**, 1564–1578.

Janbon, G., Ormerod, K.L., Paulet, D., Byrnes, E.J., Yadav, V., Chatterjee, G., Mullapudi, N., Hon, C.-C., Billmyre, R.B., Brunel, F., Bahn, Y.-S., Chen, W., Chen, Y., Chow, E.W.L., Coppée, J.-Y., Floyd-Averette, A., Gaillardin, C., Gerik, K.J., Goldberg, J., Gonzalez-Hilarion, S., Gujja, S., Hamlin, J.L., Hsueh, Y.-P., Ianiri, G., Jones, S., Kodira, C.D., Kozubowski, L., Lam, W., Marra, M., Mesner, L.D., Mieczkowski, P.A., Moyrand, F., Nielsen, K., Proux, C., Rossignol, T., Schein, J.E., Sun, S., Wollschlaeger, C., Wood, I.A., Zeng, Q., Neuvéglise, C., Newlon, C.S., Perfect, J.R., Lodge, J.K., Idnurm, A., Stajich, J.E., Kronstad, J.W., Sanyal, K., Heitman, J., Fraser, J.A., Cuomo, C.A. and Dietrich, F.S. (2014) Analysis of the genome and transcriptome of *Cryptococcus neoformans var. grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genet.* **10**, e1004261.

Janin, J. (1979) Surface and inside volumes in globular proteins. *Nature.* **277**, 491–492.

Jeffries, T.W., Grigoriev, I.V., Grimwood, J., Laplaza, J.M., Aerts, A., Salamov, A., Schmutz, J., Lindquist, E., Dehal, P., Shapiro, H., Jin, Y.-S., Passoth, V. and Richardson, P.M. (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis. Nat. Biotechnol.* **25**, 319–326.

Jones, S. and Thornton, J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology.* **272**, 121–132.

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., Davis, R.W. and Scherer, S. (2004) The diploid genome sequence of Candida albicans. *Proc. Natl. Acad. Sci. U S A.* **101**, 7329–7334.

Kamoun, S. (2006) A catalogue of the effector secretome of plant pathogenic oomycetes. *Annu Rev Phytopathol.* **44**, 41–60.

Kämper, J., Kahmann, R., Bölker, M., Ma, L.-J., Brefort, T., Saville, B.J., Banuett, F., Kronstad, J.W., Gold, S.E., Müller, O., Perlin, M.H., Wösten, H.A.B., de Vries, R., Ruiz-Herrera, J., Reynaga-Peña, C.G., Snetselaar, K., McCann, M., Pérez-Martín, J., Feldbrügge, M., Basse, C.W., Steinberg, G., Ibeas, J.I., Holloman, W., Guzman, P., Farman, M., Stajich, J.E., Sentandreu, R., González-Prieto, J.M., Kennell, J.C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Münch, K., Rössel, N., Scherer, M., Vranes, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng, S., Ho, E.C.H., Cahill, M.J., Boyce, K.J., Klose, J., Klosterman, S.J., Deelstra, H.J., Ortiz-Castellanos, L., Li, W., Sanchez-Alonso, P., Schreier, P.H., Häuser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schlüter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mannhaupt, G., Güldener, U., Münsterkötter, M., Haase, D., Oesterheld, M., Mewes, H.-W., Mauceli, E.W., DeCaprio, D., Wade, C.M., Butler, J., Young, S., Jaffe, D.B., Calvo, S., Nusbaum, C., Galagan, J. and Birren, B.W. (2006) Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis. Nature*, **444**, 97–101.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J. and Vivarès, C.P. (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature*, **414**, 450–453.

Kettles, G.J., Bayon, C., Canning, G., Rudd, J.J. and Kanyuka, K. (2017) Apoplastic recognition of multiple candidate effectors from the wheat pathogen Zymoseptoria tritici in the nonhost plant Nicotiana benthamiana. *New Phytol.* **213**, 338–350.

Kleemann, J., Rincon-Rivera, L.J., Takahara, H., Neumann, U., Ver Loren van Themaat, E., van Themaat, E.V.L., van der Does, H.C., Hacquard, S., Stüber, K., Will, I., Schmalenbach, W., Schmelzer, E. and O'Connell, R.J. (2012) Sequential delivery of host-induced virulence effectors by appressoria and intracellular hyphae of the phytopathogen *Colletotrichum higginsianum. PLoS Pathog.* **8**, e1002643.

Klosterman, S.J., Subbarao, K.V., Kang, S., Veronese, P., Gold, S.E., Thomma, B.P.H.J., Chen, Z., Henrissat, B., Lee, Y.-H., Park, J., Garcia-Pedrajas, M.D., Barbara, D.J., Anchieta, A., de Jonge, R., Santhanam, P., Maruthachalam, K., Atallah, Z., Amyotte, S.G., Paz, Z., Inderbitzin, P., Hayes, R.J., Heiman, D.I., Young, S., Zeng, Q., Engels, R., Galagan, J., Cuomo, C.A., Dobinson, K.F. and Ma, L.-J. (2011) Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathog.* **7**, e1002137.

Kohler, A., Kuo, A., Nagy, L.G., Morin, E., Barry, K.W., Buscot, F., Canbäck, B., Choi, C., Cichocki, N., Clum, A., Colpaert, J., Copeland, A., Costa, M.D., Doré, J., Floudas, D., Gay, G., Girlanda, M., Henrissat, B., Herrmann, S., Hess, J., Högberg, N., Johansson, T., Khouja, H.-R., LaButti, K., Lahrmann, U., Levasseur, A., Lindquist, E.A., Lipzen, A., Marmeisse, R., Martino, E., Murat, C., Ngan, C.Y., Nehls, U., Plett, J.M., Pringle, A., Ohm, R.A., Perotto, S., Peter, M., Riley, R., Rineau, F., Ruytinx, J., Salamov, A., Shah, F., Sun, H., Tarkka, M., Tritt, A., Veneault-Fourrey, C., Zuccaro, A., Tunlid, A., Grigoriev, I.V., Hibbett, D.S. and Martin, F. (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat. Genet.* **47**, 410–415.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

Kubicek, C.P., Starr, T.L. and Glass, N.L. (2014) Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. *Annu. Rev. Phytopathol.* **52**, 427–451.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

Laurie, J.D., Ali, S., Linning, R., Mannhaupt, G., Wong, P., Güldener, U., Münsterkötter, M., Moore, R., Kahmann, R., Bakkeren, G. and Schirawski, J. (2012) Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. *Plant Cell.* **24**, 1733–1745.

Lo Presti, L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., Zuccaro, A., Reissmann, S. and Kahmann, R. (2015) Fungal effectors and plant susceptibility. *Annu. Rev. Plant. Biol.* **66**, 513–545.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J.M., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M. and Hyman, R.W. (2005) The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science*, **307**, 1321–1324.

Lorenz, S., Guenther, M., Grumaz, C., Rupp, S., Zibek, S. and Sohn, K. (2014) Genome Sequence of the Basidiomycetous Fungus Pseudozyma aphidis DSM70725, an Efficient Producer of Biosurfactant Mannosylerythritol Lipids. *Genome Announc.* **2**,

Ma, L.-J., van der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P.M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S.E., Bluhm, B.H., Breakspear, A., Brown, D.W., Butchko, R.A.E., Chapman, S., Coulson, R., Coutinho, P.M., Danchin, E.G.J., Diener, A., Gale, L.R., Gardiner, D.M., Goff, S., Hammond-Kosack, K.E., Hilburn, K., Hua-Van, A., Jonkers, W., Kazan, K., Kodira, C.D., Koehrsen, M., Kumar, L., Lee, Y.-H., Li, L., Manners, J.M., Miranda-Saavedra, D., Mukherjee, M., Park, G., Park, J., Park, S.-Y., Proctor, R.H., Regev, A., Ruiz-Roldan, M.C., Sain, D., Sakthikumar, S., Sykes, S., Schwartz, D.C., Turgeon, B.G., Wapinski, I., Yoder, O., Young, S., Zeng, Q., Zhou, S., Galagan, J., Cuomo, C.A., Kistler, H.C. and Rep, M. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*, **464**, 367–373.

Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K.-I., Arima, T., Akita, O., Kashiwagi, Y., Abe, K., Gomi, K., Horiuchi, H., Kitamoto, K., Kobayashi, T., Takeuchi, M., Denning, D.W., Galagan, J.E., Nierman, W.C., Yu, J., Archer, D.B., Bennett, J.W., Bhatnagar, D., Cleveland, T.E., Fedorova, N.D., Gotoh, O., Horikawa, H., Hosoyama, A., Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P.R., Kato, M., Kato, Y., Kin, T., Kokubun, A., Maeda, H., Maeyama, N., Maruyama, J-I., Nagasaki, H., Nakajima, T., Oda, K., Okada, K., Paulsen, I., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K., Terabayashi, Y., Wortman, J.R., Yamada, O., Yamagata, Y., Anazawa, H., Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka, A., Isono, K., Kuhara, S., Ogasawara, N. and Kikuchi, H. (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*, **438**, 1157–1161.

Manning, V.A., Pandelova, I., Dhillon, B., Wilhelm, L.J., Goodwin, S.B., Berlin, A.M., Figueroa, M., Freitag, M., Hane, J.K., Henrissat, B., Holman, W.H., Kodira, C.D., Martin, J., Oliver, R.P., Robbertse, B., Schackwitz, W., Schwartz, D.C., Spatafora, J.W., Turgeon, B.G., Yandava, C., Young, S., Zhou, S., Zeng, Q., Grigoriev, I.V., Ma, L.-J. and Ciuffetti, L.M. (2013) Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3 (Bethesda).* **3**, 41–63.

Martin, F., Aerts, A., Ahrén, D., Brun, A., Danchin, E.G.J., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., Salamov, A., Shapiro, H.J., Wuyts, J., Blaudez, D., Buée, M., Brokstein, P., Canbäck, B., Cohen, D., Courty, P.E., Coutinho, P.M., Delaruelle, C., Detter, J.C., Deveau, A., DiFazio, S., Duplessis, S., Fraissinet-Tachet, L., Lucic, E., Frey-Klett, P., Fourrey, C., Feussner, I., Gay, G., Grimwood, J., Hoegger, P.J., Jain, P., Kilaru, S., Labbé, J., Lin, Y.C., Legué, V., Le Tacon, F., Marmeisse, R., Melayah, D., Montanini, B., Muratet, M., Nehls, U., Niculita-Hirzel, H., Oudot-Le Secq, M.P., Peter, M., Quesneville, H., Rajashekar, B., Reich, M., Rouhier, N., Schmutz, J., Yin, T., Chalot, M., Henrissat, B., Kües, U., Lucas, S., Van de Peer, Y., Podila, G.K., Polle, A., Pukkila, P.J., Richardson, P.M., Rouzé, P., Sanders, I.R., Stajich, J.E., Tunlid, A., Tuskan, G. and Grigoriev, I.V. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature*, **452**, 88–92.

Mesarich, C.H., Ökmen, B., Rovenich, H., Griffiths, S.A., Wang, C., Karimi Jashni, M., Mihajlovski, A., Collemare, J., Hunziker, L., Deng, C. (2017) Specific hypersensitive response-associated recognition of new apoplastic effectors from Cladosporium fulvum in wild tomato. *Mol. Plant. Microbe. Interact.*

Morin, E., Kohler, A., Baker, A.R., Foulongne-Oriol, M., Lombard, V., Nagy, L.G., Ohm, R.A., Patyshakuliyeva, A., Brun, A., Aerts, A.L., Bailey, A.M., Billette, C., Coutinho, P.M., Deakin, G., Doddapaneni, H., Floudas, D., Grimwood, J., Hildén, K., Kües, U., Labutti, K.M., Lapidus, A., Lindquist, E.A., Lucas, S.M., Murat, C., Riley, R.W., Salamov, A.A., Schmutz, J., Subramanian, V., Wösten, H.A.B., Xu, J., Eastwood, D.C., Foster, G.D., Sonnenberg, A.S.M., Cullen, D., de Vries, R.P., Lundell, T., Hibbett, D.S., Henrissat, B., Burton, K.S., Kerrigan, R.W., Challen, M.P., Grigoriev, I.V. and Martin, F. (2012) Genome sequence of the button mushroom *Agaricus bisporus* reveals mechanisms governing adaptation to a humic-rich ecological niche. *Proc. Natl. Acad. Sci. U S A.* **109**, 17501–17506.

Morita, T., Koike, H., Koyama, Y., Hagiwara, H., Ito, E., Fukuoka, T., Imura, T., Machida, M. and Kitamoto, D. (2013) Genome Sequence of the Basidiomycetous Yeast Pseudozyma antarctica T-34, a Producer of the Glycolipid Biosurfactants Mannosylerythritol Lipids. *Genome Announc.* **1**, e0006413.

Mosquera, G., Giraldo, M.C., Khang, C.H., Coughlan, S. and Valent, B. (2009) Interaction transcriptome analysis identifies *Magnaporthe oryzae* BAS1–4 as Biotrophy-associated secreted proteins in rice blast disease. *Plant Cell.* **21**, 1273–1290.

Nemri, A., Saunders, D.G.O., Anderson, C., Upadhyaya, N., Win, J., Lawrence, G.J., Jones, D.A., Kamoun, S., Ellis, J.G. and Dodds, P.N. (2014) The genome sequence and effector complement of the flax rust pathogen *Melampsora lini. Front Plant Sci.* **5**, 98.

O'Connell, R.J., Thon, M.R., Hacquard, S., Amyotte, S.G., Kleemann, J., Torres, M.F., Damm, U., Buiate, E.A., Epstein, L., Alkan, N., Altmüller, J., Alvarado-Balderrama, L., Bauser, C.A., Becker, C., Birren, B.W., Chen, Z., Choi, J., Crouch, J.A., Duvick, J.P., Farman, M.A., Gan, P., Heiman, D., Henrissat, B., Howard, R.J., Kabbage, M., Koch, C., Kracher, B., Kubo, Y., Law, A.D., Lebrun, M.-H., Lee, Y.-H., Miyara, I., Moore, N., Neumann, U., Nordström, K., Panaccione, D.G., Panstruga, R., Place, M., Proctor, R.H., Prusky, D., Rech, G., Reinhardt, R., Rollins, J.A., Rounsley, S., Schardl, C.L., Schwartz, D.C., Shenoy, N., Shirasu, K., Sikhakolli, U.R., Stüber, K., Sukno, S.A., Sweigard, J.A., Takano, Y., Takahara, H., Trail, F., van der Does, H.C., Voll, L.M., Will, I., Young, S., Zeng, Q., Zhang, J., Zhou, S., Dickman, M.B., Schulze-Lefert, P., Ver Loren van Themaat, E., Ma, L.-J. and Vaillancourt, L.J. (2012) Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat. Genet.* **44**, 1060–1065.

Ohm, R.A., Riley, R., Salamov, A., Min, B., Choi, I.G. and Grigoriev, I.V. (2014) Genomics of wood-degrading fungi. *Fungal Genet. Biol.* **72**, 82–90.

Panwar, V., McCallum, B. and Bakkeren, G. (2013) Endogenous silencing of Puccinia triticina pathogenicity genes through in planta-expressed sequences leads to the suppression of rust diseases on wheat. *Plant J.* **73**, 521–532.

Pedersen, C., Ver Loren van Themaat, E., McGuffin, L.J., Abbott, J.C., Burgis, T.A., Barton, G., Bindschedler, L.V., Lu, X., Maekawa, T., Wessling, R., Cramer, R., Thordal-Christensen, H., Panstruga, R. and Spanu, P.D. (2012) Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics.* **13**, 694.

Penselin, D., Münsterkötter, M., Kirsten, S., Felder, M., Taudien, S., Platzer, M., Ashelford, K., Paskiewicz, K.H., Harrison, R.J., Hughes, D.J., Wolf, T., Shelest, E., Graap, J., Hoffmann, J., Wenzel, C., Wöltje, N., King, K.M., Fitt, B.D.L., Güldener, U., Avrova, A. and Knogge, W. (2016) Comparative genomics to explore phylogenetic relationship, cryptic sexual potential and host specificity of Rhynchosporium species on grasses. *BMC Genomics.* **17**, 953.

Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods.* **8**, 785–786.

Petre, B., Joly, D.L. and Duplessis, S. (2014) Effector proteins of rust fungi. *Front Plant Sci.* **5**, 416.

Petre, B., Saunders, D.G., Sklenar, J., Lorrain, C., Win, J., Duplessis, S. and Kamoun, S. (2015) Candidate Effector Proteins of the Rust Pathogen Melampsora larici-populina Target Diverse Plant Cell Compartments. *Mol. Plant. Microbe Interact.* **28**, 689–700.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277.

Rosenblum, E.B., Fisher, M.C., James, T.Y., Stajich, J.E., Longcore, J.E., Gentry, L.R. and Poorten, T.J. (2010) A molecular perspective: biology of the emerging pathogen Batrachochytrium dendrobatidis. *Dis. Aquat. Organ.* **92**, 131–147.

Rouxel, T., Grandaubert, J., Hane, J.K., Hoede, C., van de Wouw, A.P., Couloux, A., Dominguez, V., Anthouard, V., Bally, P., Bourras, S., Cozijnsen, A.J., Ciuffetti, L.M., Degrave, A., Dilmaghani, A., Duret, L., Fudal, I., Goodwin, S.B., Gout, L., Glaser, N., Linglin, J., Kema, G.H.J., Lapalu, N., Lawrence, C.B., May, K., Meyer, M., Ollivier, B., Poulain, J., Schoch, C.L., Simon, A., Spatafora, J.W., Stachowiak, A., Turgeon, B.G., Tyler, B.M., Vincent, D., Weissenbach, J., Amselem, J., Quesneville, H., Oliver, R.P., Wincker, P., Balesdent, M.-H. and Howlett, B.J. (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat. Commun.* **2**, 202.

Rovenich, H., Boshoven, J.C. and Thomma, B.P. (2014) Filamentous pathogen effector functions: of pathogens, hosts and microbiomes. *Curr. Opin. Plant Biol.* **20**, 96–103.

Salcedo, A., Rutter, W., Wang, S., Akhunova, A., Bolus, S., Chao, S., Anderson, N., De Soto, M.F., Rouse, M., Szabo, L., Bowden, R.L., Dubcovsky, J. and Akhunov, E. (2017) Variation in the AvrSr35 gene determines Sr35 resistance against wheat stem rust race Ug99. *Science*, **358**, 1604–1606.

Sella, L., Gazzetti, K., Faoro, F., Odorizzi, S., D'Ovidio, R., Schafer, W. and Favaron, F. (2013) A *Fusarium graminearum* xylanase expressed during wheat infection is a necrotizing factor but is not essential for virulence. *Plant. Physiol. Biochem.* **64**, 1–10.

Spanu, P.D., Abbott, J.C., Amselem, J., Burgis, T.A., Soanes, D.M., Stüber, K., Ver Loren van Themaat, E., Brown, J.K.M., Butcher, S.A., Gurr, S.J., Lebrun, M.-H., Ridout, C.J., Schulze-Lefert, P., Talbot, N.J., Ahmadinejad, N., Ametz, C., Barton, G.R., Benjdia, M., Bidzinski, P., Bindschedler, L.V., Both, M., Brewer, M.T., Cadle-Davidson, L., Cadle-Davidson, M.M., Collemare, J., Cramer, R., Frenkel, O., Godfrey, D., Harriman, J., Hoede, C., King, B.C., Klages, S., Kleemann, J., Knoll, D., Koti, P.S., Kreplak, J., López-Ruiz, F.J., Lu, X., Maekawa, T., Mahanil, S.,

Micali, C., Milgroom, M.G., Montana, G., Noir, S., O'Connell, R.J., Oberhaensli, S., Parlange, F., Pedersen, C., Quesneville, H., Reinhardt, R., Rott, M., Sacristán, S., Schmidt, S.M., Schön, M., Skamnioti, P., Sommer, H., Stephens, A., Takahara, H., Thordal-Christensen, H., Vigouroux, M., Wessling, R., Wicker, T. and Panstruga, R. (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, **330**, 1543–1546.

Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D.M., Singh, K.B., Dodds, P.N. and Taylor, J.M. (2017a) LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific Reports.* **7**, 44598.

Sperschneider, J., Dodds, P.N., Gardiner, D.M., Manners, J.M., Singh, K.B. and Taylor, J.M. (2015a) Advances and challenges in computational prediction of effectors from plant pathogenic fungi. *PLoS Pathog.* **11**, e1004806.

Sperschneider, J., Dodds, P.N., Singh, K.B. and Taylor, J.M. (2017b) ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol.*

Sperschneider, J., Gardiner, D.M., Dodds, P.N., Tini, F., Covarelli, L., Singh, K.B., Manners, J.M. and Taylor, J.M. (2016) EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol.* **210**, 743–761.

Sperschneider, J., Williams, A.H., Hane, J.K., Singh, K.B. and Taylor, J.M. (2015b) Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. *Front Plant Sci.* **6**, 1168.

Stajich, J.E., Wilke, S.K., Ahrén, D., Au, C.H., Birren, B.W., Borodovsky, M., Burns, C., Canbäck, B., Casselton, L.A., Cheng, C.K., Deng, J., Dietrich, F.S., Fargo, D.C., Farman, M.L., Gathman, A.C., Goldberg, J., Guigó, R., Hoegger, P.J., Hooker, J.B., Huggins, A., James, T.Y., Kamada, T., Kilaru, S., Kodira, C., Kües, U., Kupfer, D., Kwan, H.S., Lomsadze, A., Li, W., Lilly, W.W., Ma, L.-J., Mackey, A.J., Manning, G., Martin, F., Muraguchi, H., Natvig, D.O., Palmerini, H., Ramesh, M.A., Rehmeyer, C.J., Roe, B.A., Shenoy, N., Stanke, M., Ter-Hovhannisyan, V., Tunlid, A., Velagapudi, R., Vision, T.J., Zeng, Q., Zolan, M.E. and Pukkila, P.J. (2010) Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc. Natl. Acad. Sci. U S A.* **107**, 11889–11894.

Tollot, M., Assmann, D., Becker, C., Altmuller, J., Dutheil, J.Y., Wegner, C.E. and Kahmann, R. (2016) The WOPR Protein Ros1 Is a Master Regulator of Sporogenesis and Late Effector Gene Expression in the Maize Pathogen Ustilago maydis. *PLoS Pathog.* **12**, e1005697.

Unal, C.M. and Steinert, M. (2014) Microbial peptidyl-prolyl cis/trans isomerases (PPIases): virulence factors and potential alternative drug targets. *Microbiol. Mol. Biol. Rev.* **78**, 544–571.

Upadhyaya, N.M., Garnica, D.P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., Mago, R., Cuomo, C.A., Rathjen, J.P., Park, R.F., Ellis, J.G. and Dodds, P.N. (2015) Comparative genomics of Australian isolates of the wheat stem rust pathogen Puccinia graminis f. sp. tritici reveals extensive polymorphism in candidate effector genes. *Front Plant Sci.* **5**, 759.

Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Pedro, H., Pant, R., Sadanadan, V., Khamari, L., Billal, S., Mohanty, S. and Hammond-Kosack, K.E. (2017) PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res.* **45**, D604–D610.

Vacic, V., Uversky, V.N., Dunker, A.K. and Lonardi, S. (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics.* **8**, 211.

Vleeshouwers, V.G. and Oliver, R.P. (2014) Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. *Mol. Plant. Microbe. Interact.* **27**, 196–206.

Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J.P., Shatalina, M., Roffler, S., Ben-David, R., Doležel, J., Šimková, H., Schulze-Lefert, P., Spanu, P.D., Bruggmann, R., Amselem, J., Quesneville, H., Ver Loren van Themaat, E., Paape, T., Shimizu, K.K. and Keller, B. (2013) The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nat. Genet.* **45**, 1092–1096.

de Wit, P.J.G.M., van der Burgt, A., Ökmen, B., Stergiopoulos, I., Abd-Elsalam, K.A., Aerts, A.L., Bahkali, A.H., Beenen, H.G., Chettri, P., Cox, M.P., Datema, E., de Vries, R.P., Dhillon, B., Ganley, A.R., Griffiths, S.A., Guo, Y., Hamelin, R.C., Henrissat, B., Kabir, M.S., Jashni, M.K., Kema, G., Klaubauf, S., Lapidus, A., Levasseur, A., Lindquist, E., Mehrabi, R., Ohm, R.A., Owen, T.J., Salamov, A., Schwelm, A., Schijlen, E., Sun, H., van den Burg, H.A., van Ham, R.C.H.J., Zhang, S., Goodwin, S.B., Grigoriev, I.V., Collemare, J. and Bradshaw, R.E. (2012) The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genet.* **8**, e1003088.

Xu, J., Saunders, C.W., Hu, P., Grant, R.A., Boekhout, T., Kuramae, E.E., Kronstad, J.W., Deangelis, Y.M., Reeder, N.L., Johnstone, K.R., Leland, M., Fieno, A.M., Begley, W.M., Sun, Y., Lacey, M.P., Chaudhary, T., Keough, T., Chu, L., Sears, R., Yuan, B. and Dawson, T.L. (2007) Dandruff-associated Malassezia genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. *Proc. Natl. Acad. Sci. U S A.* **104**, 18730–18735.

Zheng, P., Xia, Y., Xiao, G., Xiong, C., Hu, X., Zhang, S., Zheng, H., Huang, Y., Zhou, Y., Wang, S., Zhao, G.-P., Liu, X., St Leger, R.J. and Wang, C. (2011) Genome sequence of the insect pathogenic fungus Cordyceps militaris, a valued traditional Chinese medicine. *Genome Biol.* **12**, R116.

Zhu, Z., Zhang, S., Liu, H., Shen, H., Lin, X., Yang, F., Zhou, Y.J., Jin, G., Ye, M., Zou, H., Zou, H. and Zhao, Z.K. (2012) A multi-omic map of the lipid-producing yeast Rhodosporidium toruloides. *Nat. Commun.* **3**, 1112.

Zimmerman, J.M., Eliezer, N. and Simha, R. (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1 The UniProt search terms used for the collection of negative test sets.

Table S2 Average performance of the 50 models in 10-fold cross-validation.

Table S3 The genomes used for the evaluation.

Table S4 Effector predictions on secretomes from 93 fungal species.

Fig. S1 One of the 10 C4.5 decision trees that discriminates between fungal effectors and secreted non-effectors from pathogen secretomes [10-fold cross-validation: sensitivity, 72.3%; false positive rate, 10.2%; precision, 70.1%; area under the curve (AUC), 0.786].

Fig. S2 One of the 10 C4.5 decision trees that discriminates between fungal effectors and secreted non-effectors from pathogen secretomes [10-fold cross-validation: sensitivity, 69.1%; false positive rate, 11.7%; precision, 66.3%; area under the curve (AUC), 0.809].

Fig. S3 Distributions of all features used in the EffectorP 2.0 model. All data points were drawn on top of the box plots as black dots. Significance between groups is shown as horizontal brackets and was assessed using *t*-tests. The lower and upper hinges correspond to the first and third quartiles and the upper (lower) whiskers extend from the hinge to the largest (smallest) value that is within 1.5 times the interquartile range of the hinge. Data beyond the end of the whiskers are outliers.